A Unified STARIMA based Model for Short-term Traffic Flow Prediction

Peibo Duan[1], Guoqiang Mao[2], Wenwei Yue[3], and Shangbo Wang[4]

***Abstract*— This paper proposes a unified spatio-temporal model on the basis of STARIMA (Space-Time AutoRegressive Integrated Moving Average) for short-term road traffic prediction. The contributions of this paper are as follows. First, we develop a physically intuitive approach to traffic prediction that captures the time-varying spatio-temporal correlation between traffic at different measurement points. The spatio-temporal correlation is affected by the road network topology, time-varying speed, and time-varying trip distribution. Distinctly different from previous black-box approaches to road traffic modeling and prediction, parameters of the proposed approach have physically intuitive meanings which make them readily amenable to suit changing road and traffic conditions. Second, unlike some existing techniques which capture the variation of spatio-temporal correlation by a complete redesign and calibration of the model, the proposed approach uses a unified model which incorporates the physical factors potentially affecting the variation of spatio-temporal correlation into a series of parameters. These parameters are relatively easy to control and adjust when road and traffic conditions change, thereby greatly reducing the computational complexity. Experiments using two set of real traffic traces demonstrate that the proposed approach has superior accuracy compared with the widely used ARIMA (AutoRegressive Integrated Moving Average) and is only marginally inferior to that obtained by constructing multiple STARIMA models for different time of the day, however with a much reduced computational and implementation complexity.**

I. INTRODUCTION

Accurate prediction of short-term traffic flow can benefit both road users and traffic management authorities. On one hand, road users can use traffic prediction to make better travel decisions, choose a faster route to reach the destination, and reduce fuel costs. On the other hand, traffic management authorities can utilize traffic prediction to improve traffic operation efficiency and apply more effective traffic control strategies to alleviate traffic congestion and improve the efficiency of road networks [1], [2], [3].

Over last two decades, extensive research has been done on short-term prediction. According to the mathematical models used in these research, the techniques of traffic flow prediction can be broadly categorized into three categories: i) Parametric model whose structure has been defined in advance, e.g., the ARIMA (AutoRegressive Integrated Moving Average) model [2], [4]. ii) Non-parametric model which can capture more subtle aspects of the data and has more degrees of freedom than parametric model, e.g., ANN (Artificial Neural Network) [5], [6], iii) Hybrid integration model which combines the parametric and non-parametric model together, e.g., Bayesian classifier together with SVR (Support Vector Regression) [3].

It is worth pointing out that, there is no theoretical evidence to suggest one method is clearly superior over other methods. However, developing models that consider spatio-temporal correlation between traffic data in analysis became in the center of interest [7], [4]. Particularly, the linear parametric model, STARIMA (Space-Time AutoRegressive Integrated Moving Average) and corresponding modified versions are widely used in recent years [1], [7], [4]. Apart from the parameters as the ones in the ARIMA model that capture the temporal correlation, another set of parameters, namely spatial weight matrix, is introduced into the STARIMA model (more details of the STARIMA model will be discussed in Section II-B) with the aim of capturing spatial correlation.

However, road traffic is not stationary and the spatio-temporal correlation is time-varying. To capture this time-varying nature of road traffic and produce more accurate traffic prediction, several STARIMA based methods develop multiple STARIMA models for use during different time periods of the day [8], [1], [3]. For instance, Wanli and Laura [8] proposed a new way to calculate the spatial order between two locations by dividing their physical distance by the average speeds in different time periods. As a result, a dynamic spatial weight matrix, as well as model parameters would be re-evaluated for different time periods. Similarly, Tao et al. [1] incorporated the concepts of dynamic spatial weights and dynamic spatial neighborhood using the dynamic spatial weight matrix in order to better capture the spatial heterogeneity and temporal nonstationarity in road traffic. Later on, Athanasios et al. [4] developed a more efficient way to evaluate the spatial weight matrix for complex road structure. More precisely, they used a graph theory based technique to manage spatial dependence between road traffic, where correlations among road traffic at different locations were determined by scores calculated according to the Pearson product-moment correlation-coefficient-based metric and stored as elements of the adjacency matrix of the corresponding graphical model of the road network.

The use of multiple models may accurately capture spatio-temporal correlation variation and improve prediction accu-

[1]Peibo Duan is with School of Computing and Communications, University of Technology, Sydney, Australia Peibo.Duan@student.uts.edu.au

[2]Guoqiang Mao is with the School of Computing and Communication, The University of Technology Sydney, Australia. He is also with Data61, CSIRO, Australia Guoqiang.Mao@uts.edu.au

[3]Wenwei Yue is with School of Information Engineering, Xi'dian University, Xi'an, China Wenwei.Yue@uts.edu.au

[4]Shangbo Wang is with School of Computing and Communications, University of Technology, Sydney, Australia Shangbo.Wang@student.uts.edu.au

racy, however, it significantly increases computational complexity, especially in the case that there may be a large amount of roads contained in a road network (see the analysis in Section II-B), reducing the flexibility of a traffic flow predictor due to the correlation variability considered in these studies is not independent of model building.

To overcome these problems, in this paper, we are motivated to develop a unified STARIMA model for short-term traffic prediction to replace traditional techniques of using multiple STARIMA models for different time of the day. The contributions of the paper are briefly summarized as follows:

- We analyze the CCF (cross-correlation function) of traffic at different detector stations, and establish the time-varying spatio-temporal correlation with time-varying lag corresponding to the maximum correlation and the speed variation.
- The time-varying lag component is previously determined using training data and then applied in the STARIMA model, where all model parameters are only needed to estimate once.
- A trade-off between complexity and prediction accuracy is guaranteed. Particularly, with a theoretical analysis, the predictor has less computational complexity than the method by constructing multiple STARIMA models. Besides, experiments using real traffic data collected from a segment of a freeway are conducted, showing that the proposed predictor has superior accuracy.

The rest of the paper is organized as follows. In Section II, we give the definition of time-varying lag component along with the way to construct the unified STARIMA model. The performance of different predictors are compared using real traffic flow data in Section III. Finally, Section IV concludes the paper.

II. Approach

In this section, we commence with the estimation of time-varying lag component, which is the key step to capture temporal variation of spatio-temporal correlation. Then we depict the unified STARIMA model embedded with time-varying lag component.

A. Time-varying Lag Component

we first explore the underlying causes for the correlation variability based on the context of the data. For example, suppose there are two points A and B in a road segment. Without loss of generality, the same set of vehicles passing A will finally reach B. Therefore, it readily follows that there is a strong spatio-temporal correlation between the traffic at A and B and intuitively, the correlation will peak at different time lags corresponding to different time periods of a day. To validate such intuition, we analyze the CCF of traffic flow data measured at two traffic detector stations (s_3 and s_6) from I-80 freeway. Fig.1 and 2 show the topology of the freeway and the CCF of the traffic flow data respectively. The CCF is calculated using (1) where u and y are the traffic flow data collected at s_3 and s_6, τ is the discrete time lags in the range of $[0, 1, 2, ..., N] \subset \mathbf{N}$, where the length of one time lag, donoted by t_{lag} corresponds to $t_{lag} = 30s$, σ_{uu} and σ_{yy} are respectively the standard deviation of u and y.

$$Corr_{uy}(\tau) = \frac{E\left[(u_t - \bar{u})(y_{t+k} - \bar{y})\right]}{\sigma_{uu}\sigma_{yy}} \tag{1}$$

A higher value of CCF indicates a stronger correlation of the traffic at both stations. As shown in Fig. 2, the correlation between traffic at stations 6 and 3 peaks at different time lags depending on the time of the day. During peak period (approximately from 6:30am - 9:00am), the correlation peaks at a time lag of 3, while during off-peak period (approximately from 19:00pm - 21:30pm), the correlation peaks at a time lag of 2. We further observe that during peak hours, the time lag corresponding to the maximum correlation is larger than that for off-peak hours.

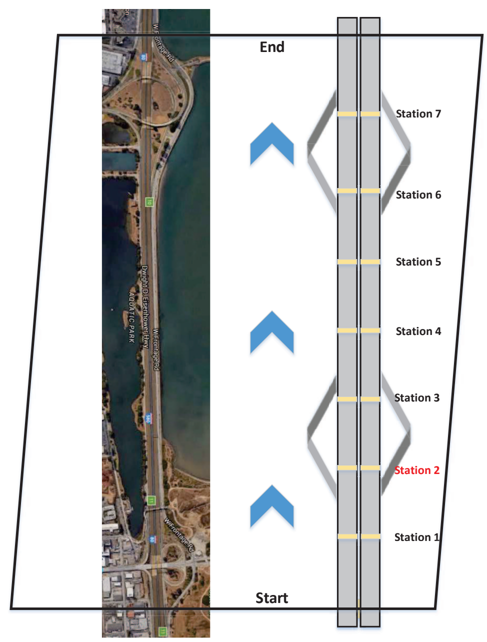

Fig. 1. The map and the topology of considered rad segment in I-80 freeway.

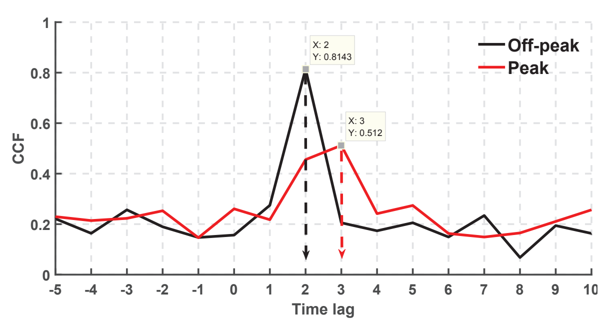

Fig. 2. The CCF between traffic flows measured at traffic detector stations 6 and 3 during two different time periods.

In practice, the difference between time lags in aforementioned two situations is related to the travel time between two locations, which is mainly attributable to the temporal speed variations. Assuming the distance between s_3 and s_6

is D and the average speed of this road segment is v in a particular time period, then the travel time between these two locations is approximately $t = D/v$. In other words, the traffic observed at s_6 at a certain time instant T is strongly correlated with that at s_3 at $T - t$. So, the time lag with the maximum correlation between s_3 and s_6 should be $\tau = [t/t_{lag}]$ where $[x]$ operator rounds x to the nearest integer.

It is worth noting that, the speed v actually refers to the spatial mean speed (SMS), which is the spatial average speed derived from the average travel time of vehicles to traverse a road segment. However, the speed collected by the dual loop detectors are the temporal mean speed (TMS) at a particular spatial location [9], [10]. In order to infer the SMS from TMS, a simple way is to regard the harmonic average of TMS as an approximate estimation of SMS [11]. Nevertheless, this method is based on an assumption that all the vehicles in each lane in one time lag have the same speed so that the TMS and the instantaneous speed is transferable. This assumption is effective when τ is small, but impractical when t_{lag} is large like 30s or more. In this case, we prefer to use a more common model which can better reflect the relationship between TMS and SMS [12]:

$$v_{tms} = v_{sms} + \frac{\sigma^2}{v_{sms}} \tag{2}$$

In (2), v_{tms} and v_{sms} are respectively TMS and SMS, and σ^2 equals to $E((v_{ins} - v_{sms})^2)$ in which v_{ins} represents instantaneous vehicle speed and $E(v_{ins}) = v_{tms}$. Then, the solution of (2), v_{sms}, can be obtained as follows [9]:

$$v_{sms} = \frac{3v_{tms} + \sqrt{9v_{tms}^2 - 8E(v_{ins}^2)}}{4} \tag{3}$$

Jiang et al. [9] assumed a quadratic relationship between $E(v_{ins}^2)$ and $E(v_{ins})$: $E[v_{ins}^2] = aE(v_{ins})^2 + bE(v_{ins}) + c$ where the parameters $\{a, b, c\}$ was estimated using 9304 samples as $\{a, b, c\} = \{1.22, -15.21, 207.95\}$.

In this paper, we use the result from [9] to obtain the SMS from the measured TMS data. Furthermore, $v_{tms} = E(v_{ins}) = \frac{\sum_{i=1}^{n} y_i v_{tms}^i}{\sum_{i=1}^{n} y_i}$ where n is the number of lanes, $y_i, i \in n$ and v_{tms}^i are respectively the traffic flow and the TMS measured by the detector located in the ith lane within a specific time period. Accordingly, the SMS between s_3 and s_6 in peak (6:30am - 9:00am) and off-peak hours (19:00pm - 21:30pm) are respectively $v_1 = 12.7251$ m/s and $v_2 = 19.1791$ m/s. Meanwhile, the time lags corresponding to the maximum CCF in these two time periods are $\tau_1 = 3$ and $\tau_2 = 2$. According to the definition of time lag τ, we have $D = v \times \tau \times t_{lag}$. Consider two speeds v_1, v_2 in two different time periods and let τ_1 and τ_2 be time lags corresponding to the maximum correlation, we are able to obtain the following equation:

$$v_1 \times \tau_1 \approx v_2 \times \tau_2 \tag{4}$$

Substituting real data into formulation (4), it is easy to find that $12.7251 \times 3 \approx 19.1791 \times 2$. This result validate our earlier conjecture in Section I and suggest that the time lag is a function of the SMS.

With the relationship between the time lags and the speed, there are two possible ways to obtain the time-varying lags between two detector stations on the one-dimensional freeway using the traffic speed measurements: 1) the time lag can be computed from the CCF of the observed traffic between two locations and its value is the time lag corresponding to the maximum CCF; 2) the time lag can be calculated using the distance between two locations divided by the average speed. Either approach has its advantages and disadvantages. For instance, more traffic measurements, e.g., speed, is required by the second approach whereas only traffic flow measurement is required in the first approach. However, the first approach requires more computations. We wish to further comment that the observed relationship between the time lag corresponding to the maximum CCF and the travel time between two locations may also provide a technique to estimate travel time from the measured traffic flow information without the need for speed measurement.

B. The Unified STARIMA Model

The STARIMA model can be expressed in the form of STARIMA(p_λ, d, q_m) where p and q are time lags for the STAR (Space-Time Autoregressive) model and the STMA (Space-Time Moving Average) model respectively, d is the degree of differencing, λ and m are the numbers of spatial lags for the STAR model and the STMA model respectively. More concretely, STARIMA(p_λ, d, q_m) is defined as follows:

$$\begin{aligned}&(I - \sum_{k=1}^{p} \sum_{l=0}^{\lambda_k} \phi_{kl} W_l L^k)(1 - L)^d Y(t) \\ =&(I - \sum_{k=1}^{q} \sum_{l=0}^{m_k} \theta_{kl} W_l L^k)\varepsilon_t.\end{aligned} \tag{5}$$

In (5), $Y(t) = \{y_1(t), y_2(t), ..., y_N(t)\}$ is a $N \times 1$ vector including the traffic flow data from N detector stations at time t; L is the lag operator by $y_i(t-1) = Ly_i(t), i \in N$; W is the spatial matrix that comprises two components: a spatial adjacency structure and a spatial weight structure. As for the spatial adjacency, it reflects first order spatial relations between all observations where two directly adjacent observations are termed as first order spatial neighbors. For the spatial weight, it reflects the spatial correlation between two first order neighbors. Generally speaking, spatial weight is either 1 or 0 where 1 means there exists correlation and 0 means the opposite.

There are three steps to set up a STARIMA model:

- **Model Identification:** using STACF (space-time autocorrelation function) and STPACF (space-time partial autocorrelation function) to identify which parameters should be estimated;
- **Parameter Estimation:** estimating the model parameters by non-linear optimization techniques;
- **Diagnostic Checking:** checking the residuals between the measurement data and that obtained from the fitted

model and determine whether an optimum model has been obtained.

The main difference between the unified STARIMA model and original one lies in that additional time-varying lags should be first considered in the phase of model identification. Given a one-dimensional freeway, the detailed process is as follows:

Step 1: Data preprocessing. Due to the existence of noise and missing in raw data collection, data cleaning is proceeded using the technique such as probabilistic principal component analysis [13]. Besides, the data is classified into two categories including training data and testing data. After filtering out the data in the festivals, the remainder is further divided into the ones in weekdays (From Mon. to Fri.) and the ones on weekends (Sat. and Sun.).

Step 2: Time periods division. According to the periodic variation of traffic flow in weekdays or on weekends, we use a clustering algorithm to allocate different time intervals during a day into a set of clusters where the label of each cluster represents a specific traffic state. For simplicity, in this paper, we roughly use two labels respectively denoted by "peak hour" and "off-peak hour". Besides, successive time intervals in a cluster consist of a time period. As the classification algorithm is not the focus of this research, we use ISODATA algorithm given in [7] which makes a classification based on traffic flow along with traffic speed.

Step 3: Time-varying lag estimation. Given a specific time period n, we first estimate the SMS between a pair of detector stations i and j, denoted by $v_{ij}(n), i, j \in N$ by means of training data. With the known topology of a road network, we then determine the distance between detector stations i and j, denoted as D_{ij}. After that, the time-varying lags between detector stations i and j, denoted by $\tau_{ij}(n)$, is calculated by $\tau_{ij}(n) = \left[\frac{D_{ij}}{v_{ij}(n)\cdot t_{lag}}\right]$. Obviously, when $i = j$, $D_{ij} = 0$ and $\tau_{ij}(n) = 0$.

Step 4: Given a specific time period, constructing the STARIMA model through model identification, parameter estimation and diagnostic checking using training data, taking into count the time-varying lags $\tau_{ij}(n)$ in different time periods n. That is, the traffic data collected at station j is time-shifted by an amount $\tau_{ij}(n)$ before the estimation of relevant model parameters.

To further illustrate how to identify parameters in our proposed unified model with consideration of time-varying spatio-temporal correlation, we take the parameter p as an example. It has been introduced earlier that p is decided by STPACF. The STPACF is calculated from Yule-Walker equation and expressed as the product of two coefficient matrices. Each element in the sequence of STPACF, $\gamma_{ij}(s)$, is denoted as the space-time covariance between the ith and jth spatial order neighbors at time lag s. It follows that $\gamma_{ij}(s)$ with time-varying lags, denoted as $\hat{\gamma}_{ij}(s + \tau_{ij}(n))$, can be calculated from (6) using training data:

$$\begin{aligned}&\hat{\gamma}_{ij}(s+\tau_{ij}(n))\\ =&\frac{1}{N(T-s)}\sum_{i=1}^{N}\sum_{t=1}^{T-s} W_i Y_i(t)(W_j L^{s+\tau_{ij}(n))} Y_i(t))\end{aligned} \tag{6}$$

Instead of spatial weight matrix W in original STARIMA model, we introduce a new $N \times N$ matrix Γ where for a spatial order $l \in \lambda$, the $(i, j)^{th}$ entry of Γ_l is $L^{\tau_{ij}(n)}$. In other words, the variation of time lag p is previously determined with time-varying lag shift. Thus, in the prediction model, the most correlated point is added by the time-varying lag. The benefit with such operation is obvious, it is not necessary to re-estimate parameters $\{p, \lambda, q, m\}$ in the STARIMA model. Instead, we only need to adjust time-varying lag in Γ.

We now analyze the computational complexity of parameters estimation for building multiple STARIMA models. According to literature [14], Dave suggested that the computational complexity of identifying parameter p using ACF (autocorrelation function) (resp. parameter q using PACF (partial autocorrelation function)) for the ARIMA model is $O(N_S N_l)$ where N_s is the number of samples from an observation and N_l is the number of time lags [14]. Unlike the ARIMA model, the parameters p and λ_ks in the STARIMA model are identified by STACF (resp. STPACF for q and m_ks). Thus, the computational complexity of calculating STACF (STPACF) between two links is $O((N - 1)N_l N_s)$ where $N - 1$ is the maximal spatial lag between two links. Consider any pair of links and the number of time periods n in a day, we have the computational complexity of parameters estimation in multiple STARIMA models is $O(nN_l N_s N^3)$.

The computational complexity of parameters estimation for the unified STARIMA model mainly relies on the following two parts: 1) the identification of time-varying lag between any pair of adjacent road segments with computational complexity $O(N)$ 2) the estimation of parameters for STARIMA model with complexity $O(N_l N_s N^3)$. As a result, the total computational complexity is $O(N + N_l N_s N^3) = O((\frac{1}{N^2} + N_l N_s)N^3)$. Generally speaking, $\frac{1}{N^2} + N_l N_s \ll nN_l N_s$. Particularly, when a large amount of road segments is considered, the proposed predictor helps to reduce computational complexity with order of $n - 1$ ($\lim\limits_{N\to\infty} \frac{1}{N^2} = 0$).

III. Experimental Validation

A. Experimental Setup

In this section, we use measurement data from a road segment on a freeway to establish the validity and accuracy of the proposed traffic flow predictor. It is part of our future work plan to extend the proposed technique to more complex urban road networks. More precisely, the data was collected from six detector stations deployed on a road segment of Interstate 80 (I-80) freeway located in Emeryville, California [15]. Furthermore, 10-days traffic data collected from April 13th to April 22th, 2005 were recorded with a sample interval of 30 seconds by means of dual-loop detectors equipped at each observation station, numbered by 1, 3, 4, 5, 6 and 7 respectively. We select the data from 8 weekdays in which the first 7-days data are used for training model and the data of the last day is used for checking the availability of the model. With the help of dual-loop detectors, the average speed in each sample interval can be calculated [16]. The speed measurement, together with known distances between stations, allows us to estimate the aforementioned time-varying lags directly. The map and the topology of considered freeway segment is shown in Fig.1. Note that traffic data at detector station 2 was not supplied. We compare our proposed technique with other two well-known approaches, viz., the multiple ARIMA/STARIMA models based approach, denoted as ARIMA*/STARIMA* in which each ARIMA/STARIMA model is set up in a specific time period of the day.

B. Experimental Results

The performance of the forecast is measured by the mean square error (MSE) and the mean absolute percentage error (MAPE). Let $\hat{y}$ be the estimate of N-dimensional vector y, the performance measure MSE can be expressed as:

$$MSE(\hat{y}, y) = \frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)^2 \tag{7}$$

MAPE is given as follows:

$$MAPE(\hat{y}, y) = \frac{1}{N}\sum_{n=1}^{N}\left|\frac{\hat{y}_n - y_n}{y_n}\right| \tag{8}$$

TABLE I

DIFFERENT TIME-VARYING LAGS DURING ONE DAY BETWEEN s_6 AND ITS lTH ($l = 1, 2, 3$) SPATIAL ORDER NEIGHBORS

Day	l= 3 (s_6, s_3)		l = 2 (s_6, s_4)		l = 1 (s_6, s_5)	
	T_1	T_2	T_1	T_2	T_1	T_2
1	3	2	2	1	1	1
2	3	2	2	1	1	1
3	3	2	2	1	1	1
4	3	2	2	1	1	1
5	3	2	2	1	1	1

To further verify the relationship between time-varying lags with speed in different time periods of one day and the accuracy of the estimates of time-varying lags used in our proposed model, which is obtained from measured speed, we use the CCF to analyze the time-varying lags over 5-days traffic flow data between station 6 and its any possible spatial order neighbors in Table I. From the table, we get knowledge that the time of one day is divided into two disjoint sets, denoted as T_1 and T_2 where $T_1 = \{(6:30am - 9:00am), (16:30pm - 18:00pm)\}$ and $T_2 = Day(24hours) - T_1$. It reveals an encouraging result that the time lags evaluated by these two methods are the same with the exception of some parts of the results in day 2. Note that the time-varying lags in T_1 and T_2 are the same when $l = 1$ between s_6 and s_5, which can be explained by the close distance between the two stations. Table I and other analysis (not shown in the paper due to space limitation) allow us to further conclude that the time-varying lag calculated between any two detector stations with different spatial orders is as those presented in formulation (9) and (10).

$$\begin{array}{c|cccc} T_1 & s_3 & s_4 & s_5 & s_6 \\ \hline s_3 & - & - & - & - \\ s_4 & 1 & - & - & - \\ s_5 & 2 & 1 & - & - \\ s_6 & 3 & 2 & 1 & - \end{array} \tag{9}$$

$$\begin{array}{c|cccc} T_2 & s_3 & s_4 & s_5 & s_6 \\ \hline s_3 & - & - & - & - \\ s_4 & 1 & - & - & - \\ s_5 & 1 & 1 & - & - \\ s_6 & 2 & 1 & 1 & - \end{array} \tag{10}$$

In Table II, we provide the MAPE/MSE of the forecast results in different time periods of the day at four detector stations using our proposed prediction technique along with STARIMA* and ARIMA*.

TABLE II

THE MAPE/MSE OF ONE-DAY TRAFFIC FLOW PREDICTION USING UNIFIED STARIMA MODEL, STARIMA*, AND ARIMA*

St.	Unified	STARIMA*	ARIMA*
s_3	17.80%/206.33	14.86%/164.21	27.33%/262.50
s_4	17.12%/191.25	15.84%/179.06	35.70%/375.64
s_5	15.13%/178.59	14.92%/159.57	29.15%/297.13
s_6	14.41%/136.27	12.65%/112.44	33.98%/342.98

From the experimental results in Table II, we can observe that the best performance is the ones obtained from STARIMA* model. Such phenomenon can be explained by that the STARIMA* based technique considers both the spatial and temporal information as well as using multiple models for different time periods for prediction. In comparison, the time-varying traffic correlation are captured using a single parameter, i.e., the time-varying lags, in the proposed technique, thereby significantly reduce computation complexity at the expense of a slight increase in prediction error ($\sim 3\%$ of the measured value). Comparing with the ARIMA* based technique, unsurprisingly, the proposed technique achieves much better prediction accuracy. Particularly, the MAPE of the proposed technique is at least 10% better than that achieved by the ARIMA* based technique.

Furthermore, Fig.3 shows the running time of each approach. More precisely, we implement different approach based on the same data at s_6 for 10 times. Each running time is the time needed for the prediction of the traffic flow in one day. From the figure, we know that the running time of our proposed unified predictor is much less than the other two methods, attributable to the single model employed for traffic prediction during different time periods.

IV. CONCLUSIONS

Motivated by the observation that spatio-temporal correlation between different detector stations is time-varying and the time lag corresponding to the maximum correlation

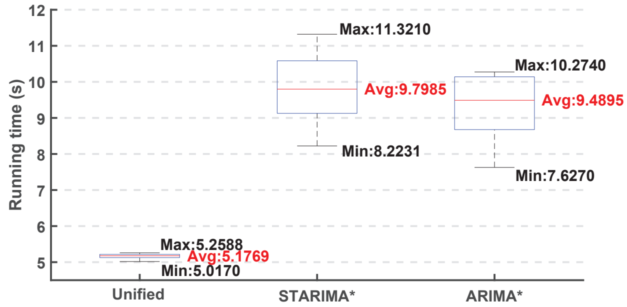

Fig. 3. The running time of the unified STARIMA model, STARIMA* and ARIMA*

approximately equals to the distance between two traffic detector stations divided by the speed of vehicles between them, in this paper, we developed a unified STARIMA model which removes the need to reconstruct different models for different time periods of the day for traffic prediction. Except a superior accuracy achieved in the case of one-dimensional freeway, the improved computational efficiency revealed through our work is that our proposed predictor can be potentially applied for a more complicated road network, e.g. the urban roads where there is a large number of road segments. However, in an urban environment, the spatio-temporal correlation between traffic tends to be much more intricate since more factors affecting such correlation should be considered like the routing selection and road topology modeling. Thus, it is part of our future work plan to extend our prediction technique to the urban roads.

References

[1] T. Cheng, J. Wang, J. Haworth, B. Heydecker, and A. Chow, "A dynamic spatial weight matrix and localized space–time autoregressive integrated moving average for network modeling," *Geographical Analysis*, vol. 46, no. 1, pp. 75–97, 2014.

[2] Y.-J. Kim, J.-s. Hong *et al.*, "Urban traffic flow prediction system using a multifactor pattern recognition model," *IEEE Trans. Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2744–2755, 2015.

[3] J. Ahn, E. Ko, and E. Y. Kim, "Highway traffic flow prediction using support vector regression and bayesian classifier," in *2016 International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2016, pp. 239–244.

[4] A. Salamanis, D. D. Kehagias, C. K. Filelis-Papadopoulos, D. Tzovaras, and G. A. Gravvanis, "Managing spatial graph dependencies in large volumes of traffic data for travel-time prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 6, pp. 1678–1687, 2016.

[5] J. Chen, K. H. Low, Y. Yao, and P. Jaillet, "Gaussian process decentralized data fusion and active sensing for spatiotemporal traffic modeling and prediction in mobility-on-demand systems," *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 3, pp. 901–921, 2015.

[6] Y.-J. Wu, F. Chen, C. Lu, B. Smith, and Y. Chen, "Traffic flow prediction for urban network using spatio-temporal random effects model," in *91st Annual Meeting of the Transportation Research Board (TRB)*, 2012.

[7] P. Duan, G. Mao, C. Zhang, and S. Wang, "Starima-based traffic prediction with time-varying lags," in *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*. IEEE, 2016, pp. 1610–1615.

[8] W. Min and L. Wynter, "Real-time road traffic prediction with spatio-temporal correlations," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 606–616, 2011.

[9] J. Han, J. W. Polak, J. Barria, and R. Krishnan, "On the estimation of space-mean-speed from inductive loop detector data," *Transportation planning and technology*, vol. 33, no. 1, pp. 91–104, 2010.

[10] J. Ahn, E. Ko, and E. Y. Kim, "Highway traffic flow prediction using support vector regression and bayesian classifier," in *2016 International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2016, pp. 239–244.

[11] R. Li, G. Rose, and M. Sarvi, "Evaluation of speed-based travel time estimation models," *Journal of transportation engineering*, vol. 132, no. 7, pp. 540–547, 2006.

[12] A. D. May, *Traffic flow fundamentals*, 1990.

[13] Y. Li, Z. Li, and L. Li, "Missing traffic data: comparison of imputation methods," *IET Intelligent Transport Systems*, vol. 8, no. 1, pp. 51–57, 2014.

[14] D. Hale, "An efficient method for computing local cross-correlations of multi-dimensional signals," *CWP Report*, vol. 656, 2006.

[15] NGSIM, "Next generation simulation," <*http://ngsim-community.org/*>, 2010.

[16] N. L. Nihan, Y. Wang, and P. Cheevarunothai, *Improving dual-loop truck (and speed) data: quick detection of malfunctioning loops and calculation of required adjustments*. Citeseer, 2006, vol. 647.