2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)
Windsor Oceanico Hotel, Rio de Janeiro, Brazil, November 1-4, 2016

# STARIMA-based Traffic Prediction with Time-varying Lags

Peibo Duan, Guoqiang Mao, Changsheng Zhang* and Shangbo Wang

*Abstract*— **Based on the observation that the correlation between observed traffic at two measurement points or traffic stations may be time-varying, attributable to the time-varying speed which subsequently causes variations in the time required to travel between the two points, in this paper, we develop a modified Space-Time Autoregressive Integrated Moving Average (STARIMA) model with time-varying lags for short-term traffic flow prediction. Particularly, the temporal lags in the modified STARIMA change with the time-varying speed at different time of the day or equivalently change with the (time-varying) time required to travel between two measurement points. Firstly, a technique is developed to evaluate the temporal lag in the STARIMA model, where the temporal lag is formulated as a function of the spatial lag (spatial distance) and the average speed. Secondly, an unsupervised classification algorithm based on ISODATA algorithm is designed to classify different time periods of the day according to the variation of the speed. The classification helps to determine the appropriate time lag to use in the STARIMA model. Finally, a STARIMA-based model with time-varying lags is developed for short-term traffic prediction. Experimental results using real traffic data show that the developed STARIMA-based model with time-varying lags has superior accuracy compared with its counterpart developed using the traditional cross-correlation function and without employing time-varying lags.**

## I. Introduction

Road traffic prediction plays an important role in intelligent transport systems by providing the required real-time information for traffic management and congestion control, as well as the long-term traffic trend for transport infrastructure planning [1]–[4]. Road traffic predictions can be broadly classified into short-term traffic predictions and long-term traffic forecasts [3], [5], [6]. Short-term prediction is essential for the development of efficient traffic management and control systems, while long-term prediction is mainly useful for road design and transport infrastructure planning.

There are two major categories of techniques for road traffic prediction: those based on non-parametric models and those based on parametric models. Non-parametric model based techniques, such as k-nearest neighbors (KNN) model [1] and Artificial Neural Networks (ANN) [7], are inherently robust and valid under very weak assumptions [8], while parametric model based techniques, such as auto-regressive integrated moving average (ARIMA) model [2], [4], [9] and its variants [10] [11], allows to integrate knowledge of the underlying traffic process in the form of traffic models that can then be used for traffic prediction. Both categories of techniques have been widely used and in this paper, we consider parametric model based techniques, particularly STARIMA (Space-Time Autoregressive Integrated Moving Average)-based techniques.

As for the estimation of parameters and coefficients in STARIMA model, overfitting easily occurs which makes the predictive performance poor as it overreacts to minor fluctuations in the training data [12]. Furthermore, the same model and hence the same correlation structure is used for traffic prediction at different time of the day, which is counter-intuitive and may not be accurate. To elaborate, consider an artificial example of two traffic stations $A$ and $B$ on a highway, where traffic station $B$ is at the down stream direction of $A$. Intuitively, the correlation between the traffic observed at $A$ and the traffic observed at $B$ will peak at a time lag corresponding to the time required to travel from $A$ to $B$ because at that time lag, the (approximately) same set of vehicles that have passed $A$ now have reached $B$. Obviously, the time required to travel from $A$ to $B$ depends on the traffic speed, which varies with the time of the day, e.g. peak hours and off-peak hours. Accordingly, the time lag corresponding to the peak correlation between the traffic at $A$ and the traffic at $B$ should also vary with time of the day and, to be more specific, should approximately equal to the distance between $A$ and $B$ divided by the mean speed of vehicles between $A$ and $B$. Therefore, in designing the STARIMA model for traffic prediction, the aforementioned time-varying lags should be taken into account for accurate traffic prediction.

To validate the aforementioned intuition, we analyze the cross-correlation function (CCF) of traffic flow data at two traffic stations (stations 6 and 3), denoted as $Corr_{63}$, from I-80 highway (more details of data are discussed in Section III-A) with the formulation (1):

$$Corr_{63} = \frac{E\left[(u_t - \bar{u})(y_{t+k} - \bar{y})\right]}{\sigma_{uu}\sigma_{yy}} \tag{1}$$

where $u$ and $y$ are the traffic flow data collected in $N$ time slots from the two traffic stations, $k$ is the temporal order in the range of $[0, 1, 2, ..., N] \subset \mathbf{N}$, $\sigma_{uu}$ and $\sigma_{yy}$ are

Asterisk is the corresponding author

Peibo Duan is with (1) the College of Computer Science and Engineering, Northeastern University, China, (2) the School of Computing and Communication, The University of Technology Sydney, Australia (e-mail: sakuragiduan@gmail.com)

Guoqiang Mao is with (1) the School of Computing and Communication, The University of Technology Sydney, Australia, (2) Data61, Sydney, Australia, (3) School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China, and (4) School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China (e-mail: g.mao@ieee.org).

Changsheng Zhang is with the College of Computer Science and Engineering, Northeastern University, China (e-mail: zhangchangsheng@neu.edu.cn)

Shangbo Wang is with the School of Computing and Communication, The University of Technology Sydney, Australia (e-mail: Shangbo.Wang@student.uts.edu.au)

978-1-5090-1889-5/16/$31.00 ©2016 IEEE

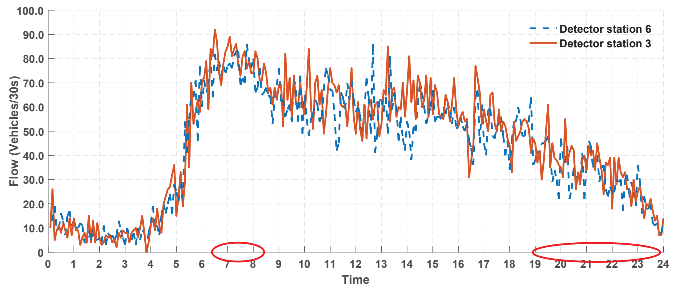

Fig. 1. Traffic flow of station 6 in one day

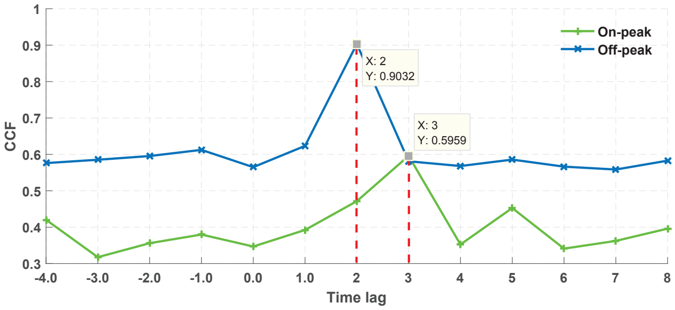

Fig. 2. The CCF between traffic stations 6 and 3 in two different time periods

respectively the standard deviation of $u$ and $y$. A higher value of CCF indicates a stronger correlation of the traffic at both stations. As shown in Fig.1, the correlation between traffic at stations 6 and 3 peaks at different time lags depending on the time of the day. During on-peak period (approximately from 6:30am - 8:30am), the correlation peaks at a time lag of 3 (one time lag corresponds to $30s$) while during off-peak period (approximately from 19pm - 24pm), the correlation peaks at a time lag of 2, where one time lag corresponds to a time of $30s$. We observe that at peak hours, the time lag corresponding to the maximum correlation is larger than that for off-peak hours. In the latter section, we will further show that this time lag approximately equals to the distance between the two traffic stations divided by the average speed. Therefore, our intuition explained in the previous paragraph is valid.

The above observation motivates us to design a STARIMA-based traffic prediction with time-varying lags which better matches the time-varying correlation structure between traffic of different stations and hence can potentially deliver more accurate traffic prediction. More specifically, the contributions of the paper are:

- We analyze the CCF between the speed and traffic flow data between different detector stations and establish the relationship between the changes in the temporal lag (corresponding to the aforementioned maximum correlation) and the speed variations.
- An unsupervised classification algorithm based on ISODATA algorithm is designed to classify different time periods of the day according to the variation of the speed. The classification helps to determine the appropriate time lag to use in the STARIMA model.
- A STARIMA-based model with time-varying lags is developed for short-term traffic prediction. Experimental results using real traffic data show that the developed STARIMA-based model with time-varying lags has superior accuracy compared with its counterpart developed using the traditional cross-correlation function and without employing time-varying lags.

The the rest of the paper is organized as follows. In Section II, we briefly discuss related work. Section III introduces the STARIMA model and the ISODATA algorithm In Section IV, we present the details the proposed algorithm. The experimental results are presented in Section V. Finally, Section VI concludes the paper.

## II. Related Work

There is previous work, which predicts traffic flow using a modified ARIMA models [4], [10], [11], [13], [14]. In [13] and [14], a multivariate ARIMA based model, ARIMAX, was applied for better traffic flow prediction. The difference is that the former paper considered the varibility of the speed from upstream to downstream, the other one considered different model specifications during different time periods of the day. Similarly, the authors in [4] also different configurations of temporal lags in ARIMA model. More concretely, they firstly applied a hidden Markov model (HMM) model along with an expectation-maximization (EM) algorithm to evaluate the traffic state (one of {Major Accident, Minor Incident, Instability, Normal Driving}) in next time slot. After that, the ARIMA models with different configurations of temporal lags were used to predict the state of the traffic flow. All these models have improved the accuracy of the forecasting results compared with ARIMA model. However, the spatial information was less considered in these models. In this way, the STARIMA based models [10], [11] have aroused more and more concern.

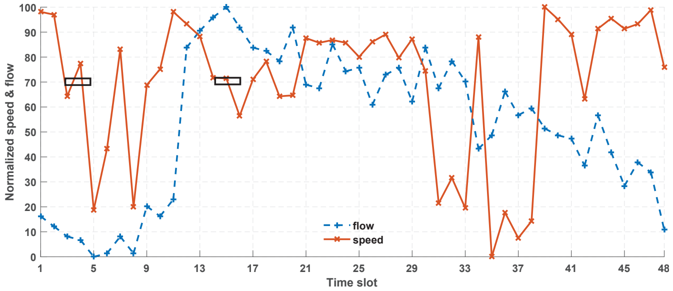

Fig. 3. The normalized freeway speed and traffic flow data in one day.

In [10], the authors proposed a dynamic STARIMA model by combining the dynamic turn ratio prediction (DTRP) model and the STARIMA model. In this paper, a dynamic space weigh matrix is used to capture different impact of traffic at upstream locations on traffic at downstream locations. Similarly, the research in [11] also applied the STARIMA model with the consideration of the dynamic space weight matrix. Our work distinguish from theirs in that in our work, the space weight matrices vary with on-peak and off-peak periods to capture the time-varying correlations of road traffic at different locations.

From the above related work, we can find that the "dynamic" of a ARIMA or STARIMA model in existing research is often used to indicate the dynamic of the space weight matrix, the traffic state during different time periods. However, sometimes it is not enough to only consider these aspects. For example, Fig.3 presents the normalized average speed and normalized flow data collected at traffic station 6 every 30 minutes in one day. Theoretically, the weight matrix in time slot 3 (or 4, the left hollow rectangle) and slot 15 (the right hollow rectangle) should be different since they are respectively in the off-peak period and peak period. However, their average speed are the same. This is caused by an inaccurate evaluation of the time range of peak or off-peak period. Furthermore, few research considered the relationship between speed and the parameters (temporal or spatial lag) in STARIMA model. Specially, a great majority of research use PACF to evaluate temporal lag which easily causes overfitting problem. Motivated by the above observations, in this paper we investigate a more efficient method to evaluate these parameters in STARIMA model with the consideration of spatial information and the variation of average speed during different time periods.

## III. Data Collection and Basic Methodologies

In this section, we briefly introduce the data we will use in the paper and the STARIMA model.

### A. Data Collection

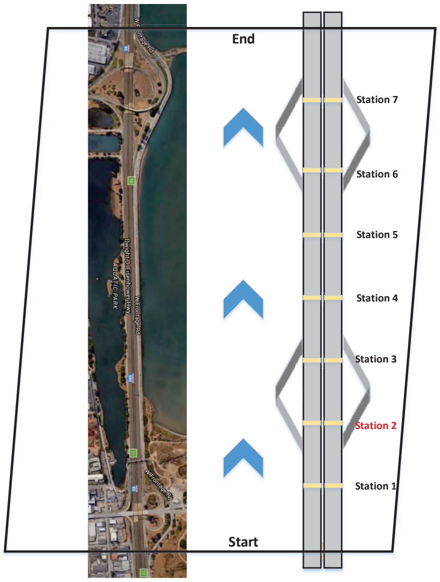

Fig. 4. The real scenario and topological structure of the I-80 freeway

In urban environment, the road structure is often complex. Also, the sensors (such as loop detectors or cameras) are not deployed at every road. Therefore, it is difficult to obtain comprehensive data. For simplicity, we only consider highway in this paper where there are only on-ramp/off-ramps so the traffic condition is comparatively simpler than that in the urban area. We use data collected from a segment of Interstate 80 (I-80) freeway located in Emeryville, California [15]. Available data are collected every 30 seconds from six traffic stations numbered by 1, 3, 4, 5, 6 and 7 within 10 days. There are two traffic stations for upstream and downstream traffic respectively. The road topology is shown in Fig.4. Note that there is no data at station 2 and there are too many interference caused by on or off-ramps at station 1 and 7. For simplicity, we only use the data collected at stations 3, 4, 5, and 6. The travel direction is from station 3 to 6.

### B. STARIMA Model

Pfeifer and Deutsch defined $STARIMA(p_\lambda, d, q_m)$ model [16] as follows:

$$(I-\sum_{k=1}^{p}\sum_{l=0}^{\lambda_k}\phi_{kl}W_lL^k)(1-L)^dY(t)= (I-\sum_{k=1}^{q}\sum_{l=0}^{m_k}\theta_{kl}W_lL^k)\varepsilon_t \tag{2}$$

where $p$ and $q$ are temporal lag for $AR$ and $MA$, $d$ is the degree of differencing. $L$ is the lag operator by which $y(t-1) = Ly(t)$. $\lambda$ and $m$ are the spatial lags for $AR$ and $MA$. Different from $y_t$ in ARIMA model,$Y(t)$ is an $N \times 1$ vector including the traffic flow data from $N$ detector stations at time $t$. In(2), $\forall l$, $W_l$ is an $N \times N$ matrix in which each non-zero element $W_l^{ij}$ reflects the correlation between location $i$ and its "$l$th order neighbors", location $j$.

There are three steps to set up a suitable STARIMA model. The first step is *Model Identification* in which the time and spatial lags are decided after an examination via PACF at spatial lag $l$ and temporal lag $k$. Furthermore, the coefficients are estimated through Yule-Walker equations. After that, *Parameter Estimation* is performed by non-linear optimization techniques. At last, *Diagnostic Checking* is implemented in order to check the residuals from the fitted model an the statistical significance of the estimated parameters using approximate confidence intervals.

## IV. STARIMA Model with Dynamic Temporal Lag

In this section, we first present a simple way to evaluate the temporal lag in relation to speed variation. Then we propose a classification algorithm based on ISODATA by which we can respectively obtain a set of speed clusters and a set of time period clusters. Finally, we describe a modified STARIMA model with the varying temporal lag.

### A. Temporal Lag with Variation of Speed

From Fig.2, the temporal lags with maximal CCF between stations 6 and 3 are different during peak and off-peak periods. This is attributable to the variation of speed. Assuming the distance between two detector stations $A$ and $B$ is $L$, and the vehicles keep a stable average speed $\bar{v}$, then approximately $t = L/\bar{v}$ is needed for vehicles to travel from $B$ to $A$. In other words, the traffic flow collected at station $A$ is strongly correlated with that at $B$ $t$ time ago. Thus the temporal lag with the maximum correlation should be $p = [t/\tau]$ where $\tau$ is the length of one temporal lag. Note that $L$ depend on the spatial order $l$ in the STARIMA model and $\bar{v}$ can often be measured by loop detectors [17]. Furthermore, the advance in telecommunication and electronic technology also brings a number of new techniques that allows us to estimate the travel time, e.g. via smartphones. Indeed, the observation discussed in the Introduction section suggests another novel way to estimate travel time: we can infer travel time from the correlation of the observed traffic.

In order to validate the above discussion, we further analyze the results in Fig.2 by using the f average speed information at stations 3 and 6, which is collected in the same day as the traffic flow data used in Fig.2. Specifically, the average speed from station 3 to station 6 between 6:30 am and 8:30 am is 44.45 feet/second. The average speed is 67.05 feet/second between 19 pm and 24 pm. The maximal temporal lag during these two time periods is respectively 3 and 2 with 30 seconds in each temporal lag. As $L = v \times p \times \tau$, given $\bar{v}_1$and $\bar{v}_2$ during two time periods along with the corresponding best temporal lag $p_1$ and $p_2$, we are able to obtain the following equation according to the theoretical analysis above:

$$v_1 \times p_1 = v_2 \times p_2 \tag{3}$$

Substituting the data into formulation (3), it is easy to find $44.45 \times 3 \approx 67.05 \times 2$. This result agrees with our theoretical analysis and verifies our speculation that temporal $p$ is a function of the variation of average speed $\bar{v}$ in Section I.

### *B. The Classification of Speed Data*

An easy way to classify speed data is by dividing into peak time and off-peak periods. After that, the temporal lag can be calculated using $p = L/\bar{v}(\pi)$, where $\bar{v}(\pi)$ is the average speed in time period $\pi, \pi \in$\{peak, Off-peak\}. However an empirical classification is often prone to error and inaccuracy. Recall the analysis in Section II, the evaluation of the average speed is sensitive to the time range selected for peak or off-peak period. It is obvious that the speed is not always fast even during off-peak period from Fig. 3. Therefore, in this paper an ISODATA[1] based speed data classification algorithm is developed to deliver an accurate classification. Using this algorithm, we firstly classify the speed data collected in each time slot into different clusters. After that, the time period clusters are confirmed based on the time slots contained in different speed clusters.

Assuming there is a set of speed data $\mathbf{v} = \{v_{t_1}, v_{t_2}, ..., v_{t_n}\}$ in which $v_{t_i}$ is the speed in time slot $t_i$. The purpose here is to confirm a set of speed clusters, denoted as $\mathit{\Gamma} = \{\mathit{\Gamma}_{v_1}, \mathit{\Gamma}_{v_2}, ..., \mathit{\Gamma}_{v_N}\}$, where $\forall \mathit{\Gamma}_{v_{i \in N}} \subset \mathbf{v}$ with cluster center $v_i$ and $\forall i, j \in N, \mathit{\Gamma}_{v_i} \cap \mathit{\Gamma}_{v_j} = \emptyset$. Based on $\mathit{\Gamma}$, we can obtain another set of time period clusters, denoted as $\Omega = \{\Omega_1, \Omega_2, ..., \Omega_{|\mathit{\Gamma}|}\}$, in which $\forall \Omega_i = \{T_i^1, T_i^2, ..., T_i^{K_i}\}$.

**Algorithm 1** Speed Data Classification

```
 1: Input: $\boldsymbol{K_{max}}$, $\boldsymbol{n_{min}}$,$\boldsymbol{\sigma^2_{max}}$,$\boldsymbol{d_{min}}$,$\boldsymbol{I}$,$\boldsymbol{v}$,$\Delta$
 2: Return: $\boldsymbol{\mathit{\Gamma}}$,$\boldsymbol{\Omega}$
 3: $\boldsymbol{\mathit{\Gamma}}$←ISODATA($\boldsymbol{K_{max}}$,        $\boldsymbol{n_{min}}$,$\boldsymbol{\sigma^2_{max}}$,$\boldsymbol{d_{min}}$,$\boldsymbol{L}$,$\boldsymbol{v}$)

 4: for  $\forall \boldsymbol{\mathit{\Gamma}}_{v_i} \in \boldsymbol{\mathit{\Gamma}}$  do
 5:     $\forall \boldsymbol{\Omega}_i = \{T_i^1, T_i^2, ..., T_i^{K_i}\}, \emptyset \rightarrow \forall T_i^k \in \boldsymbol{\Omega}_i$
 6:     $\forall T_i^1 \leftarrow t_1, v_{t_1} \in \boldsymbol{\mathit{\Gamma}}_{v_i}$
 7:     for  $\forall v_{t_j} \in \boldsymbol{\mathit{\Gamma}}_{v_i}$  do
 8:         for  $\forall T_i^k \in \boldsymbol{\Omega}_i$  do
 9:             if  $\exists t \in T_i^k$ and $t \pm 1 = t_j$  then
10:                 $T_i^k \leftarrow t_j$
11:             end if
12:         end for
13:     end for
14:     for  $\forall T_i^k \in \boldsymbol{\Omega}_i$  do
15:         $m = |T_i^k|$
16:         if  $m < \Delta$  then
17:             for  $\forall t_j \in T_i^k$  do
18:                 $t_j \quad \rightarrow \quad min\{D(t_j, T_i^{\bar{k}}) | \forall \bar{k} \in K_i, \bar{k} \neq k\}$

19:             end for
20:             $\boldsymbol{\Omega}_i - \{T_i^k\}$
21:         end if
22:     end for
23:     $\boldsymbol{\Omega} \cup \boldsymbol{\Omega}_i$
24: end for
```

In addition, let $T_i^k, k \in K_i$ be a set of continuous time slots, termed as a time range and defined as follows:

$$T_i^k = \underbrace{\{t_j, t_j + 1, ..., t_j + m\}}_{m \geqslant \Delta} \subset \mathit{\Gamma}_{v_i} \tag{4}$$

where $m$ is the number of time slots contained in $T_i^j$, $\Delta$ is a threshold defined as the minimal number of time slots included in a time period. The speed data classification algorithm is given in Algorithm 1. In line 3, the ISODATA algorithm is implemented to get speed clusters. The time period clusters are obtained from line 4 to 23. It is worth mentioning that a decision is made to decide whether $T_i^k$ belongs to $\Omega_i$ by comparing its capacity and threshold $m$ (from line 16 to 23). If $T_i^k$ does not belong to $\Omega_i$, line 17 and 18 are implemented to allocate each $t_j \in T_i^k$ to other $T_i^{\hat{k}}, \hat{k} \neq k$ by the operation $min\{D(t_j, T_i^{\bar{k}}) | \forall \bar{k} \in K_i, \bar{k} \neq k\}$. $D(t_j, T_i^{\bar{k}})$ is defined as the absolute difference between speed recorded in time slot $t_j$ and the average speed calculated during time period $T_i^{\bar{k}}$, which is presented in (5).

$$D(t_j, T_i^{\bar{k}}) = |v_{t_j} - \frac{\sum_{t_{\bar{j}} \in T_i^{\bar{k}}} v_{t_{\bar{j}}}}{|T_i^{\bar{k}}|}| \tag{5}$$

### *C. $STARIMA(\lambda, p_\lambda(v), d, q_m)$ Model*

According to the speed and time period clusters obtained from Section IV-B, we propose a modified STARIMA model, denoted as $STARIMA(\lambda, p_\lambda(v), d, q_m)$. The definitions of

[1] More details about the process of ISODATA algorithm are available in reference [18].

parameters $\lambda$, $d$ and $q_m$ in this model are the same as the original STARIMA model, except that the temporal lag $p$ will vary with the spatial order $l$ and the average speed in different time periods. More precisely, given a time period $T_i^k \in \Omega_i$, $STARIMA(\lambda, p_\lambda(v), d, q_m)$ is defined as follows:

$$(I - \sum_{l=0}^{\lambda} \phi_l W_l (P_l(\bar{v}_i^k)L))(1-L)^d Y(t) = (I - \sum_{k=1}^{q} \sum_{l=0}^{m_k} \theta_{kl} W_l L^k)\varepsilon_t \tag{6}$$

In 6, $P_l(\bar{v}_i^k)$ is a $N \times N$ vector in which each element $p_l^{mn}(\bar{v}_i^k)$ represents the temporal lag between two station $s_m$ and $s_n$ with the spatial order $l$. $p_l^{s_1 s_2}(\bar{v}_i^k)_{ij}$ is calculated by $L(l)/\bar{v}_i^k$ where $L(l)$ is the distance between these two stations and $\bar{v}_i^k$ is the average speed in time period $T_i^k$ which is equal to $\frac{\sum_{t_j \in T_i^k} v_{t_j}}{|T_i^k|}$. Note that when $l = 0$, the "0th order neighbor" of a station is itself such that the temporal lag is evaluated with the PACF used in ARIMA.

## V. Experimental Validation

Based on the data collection introduced in Section III, we utilize the speed and traffic flow data at stations 3, 4, 5 and 6. At each station, there are 2880 data recorded in one day and the length of one time slot is 30 seconds. In order to eliminate noise in the data, we make a "smooth" operation by calculating the mean traffic flow every $x$ data points and regarding it as one data point. The experimental results are divided into two parts. In the first part, we provide the speed and time period clusters classified by our proposed algorithm. For the speed data $\mathbf{v}$, we choose $x = 30$. In the second part, we present the forecast results of traffic flow in different time periods and stations using $STARIMA(\lambda, p_\lambda(v), d, q_m)$ model. We choose $x = 4$ to calculate the mean traffic flow using original traffic flow data within 2 minutes.

### A. The Speed and Time Period Clusters

Firstly, the configuration of input parameters of algorithm is given in Table I in which the speed data $\mathbf{v}$ is the results after the smooth operation on the speed data collected from four stations. With this setting, the smallest length of time range $T_i^k \in \Omega_i$ is 120 minutes. The speed and time period clusters classified by Algorithm 3 is presented in table II.

TABLE I

The input parameters in Algorithm 3

| Parameters | Value | Parameters | Value |
|---|---|---|---|
| $K_{max}$ | 3 | $d_{min}$ | 30 |
| $n_{min}$ | 5 | $I$ | 10 |
| $\delta^2_{max}$ | 15 | $\Delta$ | 8 |

TABLE II

The speed and time period clusters

| Clusters | Values |
|---|---|
| Speed | $\mathit{\Gamma} = \{\mathit{\Gamma}_{v_1}, \mathit{\Gamma}_{v_2} \| v_1 = 82.15, v_2 = 34.33\}$ |
| Time period | $\Omega = \{\Omega_1, \Omega_2 \| \Omega_1 = \{T_1^1, T_1^2, T_1^3, T_1^4\}, \Omega_2 = \{T_2^1, T_2^2, T_2^3\}\}$<br>$T_1^1$=[0,2am), $T_1^2$=[4-6:30am),<br>$T_1^3$=[10am-15pm), $T_1^4$=[18:30-24pm]<br>$T_2^1$=[2,4am), $T_2^2$[6:30-10am),<br>$T_2^3$=(15-18:30pm] |

TABLE III

The temporal lag in $T_2^2$ and $T_1^4$ with different spatial order

| Day | $l$= 3 $(s_6, s_3)$ | | $l = 2$ $(s_6, s_4)$ | | $l = 1$ $(s_6, s_5)$ | |
|---|---|---|---|---|---|---|
| | $T_2^2$ | $T_1^4$ | $T_2^2$ | $T_1^4$ | $T_2^2$ | $T_1^4$ |
| 1 | 3/3 | 2/2 | 2/2 | 1/1 | 1/1 | 1/1 |
| 2 | **3/4** | **2/-9** | 2/2 | 1/1 | **1/-3** | 1/1 |
| 3 | 3/3 | 2/2 | 2/2 | 1/1 | 1/1 | 1/1 |
| 4 | 3/3 | 2/2 | 2/2 | 1/1 | 1/1 | 1/1 |
| 5 | 3/3 | 2/2 | 2/2 | 1/1 | 1/1 | 1/1 |

There are two speed clusters in $\mathit{\Gamma} = \{\mathit{\Gamma}_{v_1}, \mathit{\Gamma}_{v_2}\}$, in which $v_1$is the cluster center whose value is 82.15 feet/s and $v_2$ is 34.33 feet/s. Based on the speed clusters, one day is divided into 2 clusters with $\Omega = \{\Omega_1, \Omega_2\}$ in which $\Omega_1$includes 4 time ranges and $\Omega_2$ includes 3 time ranges. According to such classification, we present the temporal lag calculated by (3) (left part of "/" in each column) and by the CCF (the right part of "/" in each column) during the time range $T_2^2$(6:30-10am) and $T_1^4$(18:30-24pm) with the spatial lag $l = 1, 2, 3$. The results are shown in Table III. It reveals an encouraging result that the temporal lags evaluated by these two methods are the same with the exception of some parts of the results in day 2. In addition, comparing the temporal lags evaluated upon different spatial lags, we can find that the temporal lags during peak and off-peak periods are the same when $l = 1$ ($s_6$ and $s_5$). This is because the temporal lag is less affected by the speed if two stations are very close.

### B. Results of Traffic Flow Prediction

We choose the traffic flow data in day 3. After the smooth operation, we predict the traffic flow in one hour (30 time slots), and other data during each time range $T_i^k \in \Omega_i$ are used for training $STARIMA(\lambda, p_\lambda(v), d, q_m)$ model (for simplicity, denoted as $STARIMA(p(v))$). The settings are the same when using ARIMA and Chaos theory based model (abbreviated as Chaos) [19]. The performance of the forecast is measured by the mean square error (MSE) and the mean absolute percentage error (MAPE).

TABLE IV

The MAPE/MSE of four stations using

| $T_i^k$ | $s_3$ | $s_4$ | $s_5$ | $s_6$ |
|---|---|---|---|---|
| $T_1^1$ | 4.00%/16.74 | 1.14%/11.33 | 3.00%/29.402 | 5.43%/28.12 |
| $T_1^2$ | 11.82%/46.04 | 6.78%/63.11 | 1.54%/11.63 | 6.16%/39.41 |
| $T_1^3$ | 17.85%/87.90 | 14.34%/95.31 | 10.99%/88.97 | 12.62%/70.69 |
| $T_1^4$ | 13.69%/71.16 | 2.93%/28.33 | 3.96%/27.15 | 7.28%/44.47 |
| $T_2^1$ | 12.15%/ 58.60 | 3.88%/36.06 | 2.78%/26.39 | 7.00%/45.98 |
| $T_2^2$ | 14.00%/75.35 | 4.86%/59.51 | 2.35%/29.03 | 8.03%/51.01 |
| $T_2^3$ | 12.20%/62.79 | 6.61%/61.82 | 4.24%/52.79 | 8.68%/55.56 |

We provide the MAPE/MSE of four stations using our proposed model in Table IV. Combining with Fig. 5, we can see that the forecast results are inspiring. Especially, the MAPE of station 4, 5 and 6 are all below 9% except $T_1^3$ (10am-15pm), which is attributable to the frequent fluctuation of

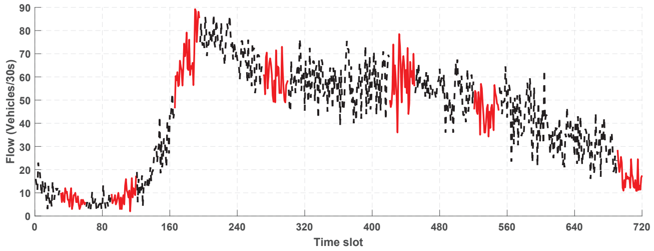

Fig. 5. The forecasting results for one day at station 6

traffic flow during this time range as shown Fig. 1. As the minimal time range based on the configuration of $\Delta = 8$ is 120 minutes, it did not capture such frequent fluctuation. In addiction, the MAPE/MSE of station 3 are higher than other stations because none neighbors are considered in this paper at this station such that the $\lambda = 0$. In this way, our model is actually similar with a ARIMA model.

TABLE V
THE MAPE/MSE OF 4 STATIONS USING $STARIMA(p(v))$, CHAOS, AND $ARIMA(2,1,2)$

| St. | $STARIMA(p(v))$ | $Chaos$ | $ARIMA(2,1,2)$ |
|---|---|---|---|
| $s_3$ | 12.25%/59.80 | 11.57%/47.94 | 34.26%/95.62 |
| $s_4$ | 5.51%/36.79 | 7.49%/66.54 | 32.56%/127.90 |
| $s_5$ | 4.02%/37.71 | 13.64%/71.79 | 25.64%/102.87 |
| $s_6$ | 7.82%/43.26 | 10.41%/56.02 | 28.66%/96.76 |

TABLE VI
THE MAPE/MSE BASED ON FORECASTING RESULTS IN 9-10AM AND 23-24PM AT STATION 6

| St. | $STARIMA(p(v))$ | $Chaos$ | $ARIMA(2,1,2)$ |
|---|---|---|---|
| 9-10am | 1.28%/6.52 | 13.01%/44.70 | 13.11%/42.74 |
| 23-24pm | 3.38%/15.55 | 6.12%/28.55 | 34.82%/98.11 |

In Table V, we compare the MAPE/MSE of one day using $STARIMA(p(v))$, Chaos and $ARIMA(2,1,2)$ at four stations. Except station 3, all the MAPE/MSE achieved by our model are better than those of the other two models. Furthermore, in Table VI, we present the MAPE/MSE of the forecast results of station 6 using these three models, in which the forecast time ranges are 9-10am and 23-24pm. It can be seen that the performance of our model is almost coincident with the true data. And Chaos comes to the second in the prediction during 23-24pm.

## VI. CONCLUSIONS

Motivated by the observation that the correlation between traffic at different traffic stations is time-varying and the time lag corresponding to the maximum correlation approximately equals to the distance between two traffic stations divided by the speed of vehicles between them, in this paper, we developed a modified STARIMA model with time-varying lags for short-term traffic flow prediction. Experimental results using real traffic data collected on a highway showed that the developed STARIMA-based model with time-varying lags has superior accuracy compared with its counterpart developed using the traditional cross-correlation function and without employing time-varying lags. In an urban environment, the correlation between traffic tends to be much more intricate. It is part of our future work plan to develop prediction technique for urban roads that incorporates the knowledge of the underlying road topology.

## REFERENCES

[1] P. Dell'Acqua, F. Bellotti, R. Berta, and A. De Gloria, "Time-aware multivariate nearest neighbor regression methods for traffic flow prediction," *IEEE Trans. Intelligent Transportation Systems*, vol. 16, no. 6, pp. 3393–3402, 2015.

[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] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-term traffic forecasting: Where we are and where we're going," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 3–19, 2014.

[4] G. Comert and A. Bezuglov, "An online change-point-based model for traffic parameter prediction," *IEEE Trans. Intelligent Transportation Systems*, vol. 14, no. 3, pp. 1360–1369, 2013.

[5] M. Lippi, M. Bertini, and P. Frasconi, "Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning," *IEEE Trans. Intelligent Transportation Systems*, vol. 14, no. 2, pp. 871–882, 2013.

[6] F. G. Habtemichael and M. Cetin, "Short-term traffic flow rate forecasting based on identifying similar traffic patterns," *Transportation Research Part C: Emerging Technologies*, 2015.

[7] L. Song, "Improved intelligent method for traffic flow prediction based on artificial neural networks and ant colony optimization." *Journal of Convergence Information Technology*, vol. 7, no. 8, 2012.

[8] 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.

[9] D. Billings and J.-S. Yang, "Application of the arima models to urban roadway travel time prediction-a case study," in *2006 IEEE Int. Conf. Systems, Man and Cybernetics*, vol. 3, pp. 2529–2534.

[10] X. Min, J. Hu, Q. Chen, T. Zhang, and Y. Zhang, "Short-term traffic flow forecasting of urban network based on dynamic starima model," in *2009 IEEE Int. Conf. Intelligent Transportation Systems*, pp. 1–6.

[11] 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.

[12] B. Everitt, *The Cambridge dictionary of statistics/BS Everitt*. Cambridge University Press, Cambridge, UK New York:, 2002.

[13] B. Williams, "Multivariate vehicular traffic flow prediction: evaluation of arimax modeling," *Journal of the Transportation Research Board*, no. 1776, pp. 194–200, 2001.

[14] A. Stathopoulos and M. G. Karlaftis, "A multivariate state space approach for urban traffic flow modeling and prediction," *Transportation Research Part C: Emerging Technologies*, vol. 11, no. 2, pp. 121–135, 2003.

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

[16] P. E. Pfeifer and S. J. Deutrch, "A three-stage iterative procedure for space-time modeling phillip," *Technometrics*, vol. 22, no. 1, pp. 35–47, 1980.

[17] Y. Wang and N. Nihan, "Freeway traffic speed estimation with single-loop outputs," *Journal of the Transportation Research Board*, no. 1727, pp. 120–126, 2000.

[18] G. H. Ball and D. J. Hall, "Isodata, a novel method of data analysis and pattern classification," DTIC Document, Tech. Rep., 1965.

[19] W.-C. Hong, "Application of seasonal svr with chaotic immune algorithm in traffic flow forecasting," *Neural Computing and Applications*, vol. 21, no. 3, pp. 583–593, 2012.