Transportation Research Part C 19 (2011) 606–616

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Real-time road traffic prediction with spatio-temporal correlations

Wanli Min [a], Laura Wynter [b,*]

[a] IBM Singapore, Changi Business Park, Singapore 486072, Singapore
[b] IBM T.J. Watson Research Center, PO Box 218, Yorktown Heights, NY 10598, United States

ARTICLE INFO

Article history:
Received 21 July 2009
Received in revised form 13 June 2010
Accepted 17 October 2010

Keywords:
Intelligent transport systems
Volume
Speed
Predictive modeling

ABSTRACT

Real-time road traffic prediction is a fundamental capability needed to make use of advanced, smart transportation technologies. Both from the point of view of network operators as well as from the point of view of travelers wishing real-time route guidance, accurate short-term traffic prediction is a necessary first step. While techniques for short-term traffic prediction have existed for some time, emerging smart transportation technologies require the traffic prediction capability to be both fast and scalable to full urban networks. We present a method that has proven to be able to meet this challenge. The method presented provides predictions of speed and volume over 5-min intervals for up to 1 h in advance.

1. Introduction

Real-time road traffic prediction is a capability that traffic authorities have begun demanding in increasing number. In the previous decade, the collection of real-time traffic data was a foremost goal. Now that many traffic authorities possess real-time traffic data feeds and information warehouses containing extensive traffic data, the most sophisticated have begun moving to the next logical step: specifically, leveraging the vast stores of data and feeds for real-time forward-looking analysis.

Road traffic prediction is the first major step in that direction. Whereas tools exist today to provide traffic control assistance as well as traveler information from real-time data, such tools are not widely available using future predictive information. However, it is clearly of interest to instruct traffic controllers on how best to set signals or variable message signs based on expected traffic conditions in the near future, rather than based on a traffic situation soon to be obsolete. Similarly, a traveler would prefer to be given route guidance information corresponding to the likely traffic when she will be on the roads in question, rather than on the condition that occurred prior to her starting her journey. The latter is particularly true for medium-length trips, such as 10-min or more in duration.

Because of increasing demand for road traffic predictive tools, the body of literature on traffic prediction methods has increased substantially in the past decade. Much of the work still focuses on expressways, although the increasing emergence of data collection on full urban networks is shifting the trend. The literature tends to focus as well on smoothed data, such as 15-min smoothed averages, and does not often predict more than a single time point into the future. However, extending beyond those two constraints to shorter time intervals and predicting on several time periods into the future allows for a wider range of applications to make use of the predictions. Thirdly, the methods used in practice for real-time traffic prediction should be fast and scalable. This paper seeks to fill precisely that gap.

We are interested here in predicting both traffic speeds and traffic volumes. Speeds are typically averaged over multiple data readings and aggregated up to a common time cadence. In our case, we have used 5-min speed readings as the real-time

* Corresponding author.
E-mail addresses: minw@sg.ibm.com (W. Min), lwynter@us.ibm.com (L. Wynter).

0968-090X/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.trc.2010.10.002

data source. Volumes also are averaged so that they represent 5-min intervals, but the values are often scaled to hourly volume levels. In both cases, multiple lanes are included in a single speed or volume reading.

Using an approach that bears some similarity to that which we propose, Smith et al. (2002) performed a comparison of a seasonal ARIMA model and a nearest neighbor technique to predict 15-min highway traffic flow rates from two loop detector locations on an expressway around London. They conclude that the seasonal ARIMA model offers better accuracy than the heuristic nearest neighbor method, at the price of more expensive computational characteristics. Although their ARIMA model outperforms the other methods they tested, its accuracy does not match that of our method on similar data. Furthermore, their ARIMA approach is highly computationally intensive, even on only two sensor locations. Another similar approach was taken by the authors in A–Deek et al. (2001) though their results on predictions at individual locations, rather "corridor" or average values over multiple locations, are not very accurate. In addition, the authors state that they are unable to predict with reasonable accuracy congested conditions. A recent reference by Chandra and Al-Deek (2009) also focuses on ARIMA models and in many ways is of a similar spirit to our work; however, those authors do not propose any particular approach for capturing network effects and at the same time avoiding over-specification, as we do in this paper.

In addition, there are a number of references in which real-time traffic prediction is performed using neural networks, such as (van Lint et al., 2005; Vlahogianni et al., 2005; Zheng et al., 2006). The reference Zheng et al. (2006) examines 15-min traffic volumes during daytime hours on an expressway, and attempts to predict the volume for the next 15-min interval using a combination of methods, including neural networks. While the accuracy presented in the last paper is reasonable, the context is somewhat limited: three points on a single expressway were considered during daytime hours on weekdays. The accuracy obtained with the method we propose in this paper is higher on analogous 15-min averaged data on the same roadway. Furthermore, no information on the computational overhead required by the neural network approach was provided, but it appears unlikely that such methods could scale to full metropolitan networks for use in real-time.

Kamarianakis and Prastacos (2003) estimate parameters in a model that takes into account both spatial and temporal correlations across the road network. While the basic form of their model has some similarity to ours, significant differences exist. Their model requires estimating a very large number of parameters, and yet does not take into account several important characteristics of a transportation network. In particular, they assume that the spatial correlations are represented by a fixed set of matrices, which depend upon the distances between links. However, on a transportation network, depending upon whether a link is congested or not, the other network links influencing its traffic flow will vary considerably. This is not captured by the approach of Kamarianakis and Prastacos (2003). Furthermore, the method proposed by the authors in Kamarianakis and Prastacos (2003) assumes stationarity of the system. While the authors note that the traffic flow parameters are clearly not stationary over the time period being modeled, they propose to perform differencing of the data points, with a differencing period of 1 day. This does not, however, deal with inter-day fluctuations, which should violate the stationarity assumption and introduce non-negligible bias into the estimated parameters. A more recent work by the authors makes use of GARCH models to handle the fact that variance in the data is different at peak and off-peak times; however, the accuracy achieved was quite poor. A line of references by Wang et al. makes use of macroscopic traffic flow modeling for real-time traffic prediction. That approach can handle only segments of expressways and while no numerical accuracy is provided by the authors, the graphics do not suggest a level of accuracy near that which is provided by our method (see Wang et al., 2007 and references therein).

In general, most references available are limited to expressways during daytime hours; our goal was to develop a traffic prediction methodology robust and accurate enough to handle the full range or urban roads as well as nighttime and weekends. Furthermore, in the literature, spatial correlations are taken into account in a limited manner, such as through incorporating the effect of a couple of upstream highway links. The drawback is that, on an urban traffic network where real-time data may be missing at some time periods, the links' data may not be available. Hence, an approach limited to using no or very few interactions between links may not be able to take into account relevant traffic elsewhere on the network. Finally, the majority of references available make use of smoothed data (for example, by considering 15-min average values) and handle predictions of one time point into the future. However, much real-time data is provided on a 5-min basis. Hence, a method should be able to react on that finer (and more volatile) time scale. In addition, predictions only one time point into the future are of limited value for certain applications whose applicability is further into the future, such as optimal traveler routing.

For these reasons, and with the goal of enabling a traffic prediction tool that can run network-wide in real-time, we have developed an extended time-series-based approach, where the extension takes into account spatial and temporal interactions in a new manner, specialized to the context of road traffic.

The next section presents the model and the network description used by the model. The following section provides numerical examples on our test network. Lastly, we present our conclusions and recommendations for further work.

2. The model

In performing traffic prediction on a road network, it is important to consider both the completeness of the model, in terms of the number and type of parameters that figure into the predictive model, and the calculability of the model. The two goals are typically at odds with each other: the first goal leads to a specification of a greater number of estimation

parameters, whereas the second goal seeks to reduce that number. Our model attempts to reconcile these two conflicting goals by leveraging the structure of a transportation network.

Furthermore, we are interested in developing a model that can be used for a full urban network in real-time, with a high degree of accuracy on 5-min speed and volume readings. Our initial focus in the model has been on speed and volume predictions, though the model has been tested successfully on other traffic metrics available in real-time, such as occupancy and link travel time. An interesting question is whether or not the model for a particular traffic parameter should leverage input data on other potentially available traffic parameters. In theory, one would suspect that the answer is a resounding "yes", an in particular when speed and volume are available, since it is well known that the speed-flow relationship is bi-valued, and hence knowing the traffic speed should be of use in determining which regime the volume is in. However, in practice the benefit of a statistical model of both volume and speed is not as valuable as it appears. In fact, in a purely empirical observation Chandra and Al-Deek developed one of their many models with both speed and volume terms and concluded that there was no benefit with respect to the either single-parameter-type model. Indeed, in practice, the data used to generate speed and volume estimates in real-time are (i) heterogeneous in terms of how they are produced, and hence in their statistical properties, and (ii) not necessarily available for the same time step on the same link at the same time.

Our approach must satisfy certain strong constraints including availability, scalability, and robustness of the solution. Hence, given the realities of real-time traffic data, a conscious choice was made not to combine different types of traffic data in a single link-level model.

A second conceptually valuable source of information for real-time road traffic prediction is incident, including extreme weather, data. Similar to the above discussion, one would expect that incident data should improve the predictive quality of a statistical model. While, in some cases, that may be the case, and the same for weather data, in our in-depth research into real-time traffic data, the predictive quality of the incident data was found to be very low. Again, some knowledge of the source of the data is helpful in explaining why it is so. In cases of incident data that we explored, there appeared to be a one-time-step time lag in the entry of that incident data code. In other words, it was possible to observe the shock to the traffic parameter (speed or volume) in the time step preceding the time step at which the incident was first registered. Hence, our models are able to detect the incident before the code is issued. This is due to the way in which traffic and incident data are aggregated by the traffic authority, and may, of course, vary from one implementation to another.

2.1. Notation and basic relations

Let j be the location index, t as time-of-day index, and r be the *template* index to be described further later. The overall model structure is

$$y_{jtr} = \mu_{jtr} + x_{jtr} \tag{1}$$

where μ_{jtr} is the time and space-dependent mean value. We propose obtaining μ by some form of weighted average. The precise form and the weights should be calibrated to best reflect the traffic, and should be re-calibrated periodically. The term x_{jtr} denotes the deviation from the mean; this transient model is of critical importance to short-term predictions.

The time and space-dependent mean value reflects the typical behavior of traffic at a finely granular level and permits the use of a separate transient model to capture variations. The choice of a method for defining the time and space-dependent mean value is hence important to the success of the overall model.

We make use of a form of weighted average which weights more heavily the recent past to the more distant past.

2.2. The transient model

2.2.1. Basic relations of the transient model

We adopt a multivariate spatial-temporal autoregressive (MSTAR) model to account for transient behavior on the traffic network. The standard Vector-ARMA(p,q), or VARMA(p,q), model is:

$$\left[I - \sum_{d=1}^{p} \Phi_d B^d\right] \boldsymbol{X}_t = \left[I + \sum_{d=1}^{q} \Theta_d B^d\right] \boldsymbol{a}_t \tag{2}$$

where B is a back-shift operator, so that

$$B^d \boldsymbol{X}_t = \boldsymbol{X}_{t-d}$$

The parameters Φ represent the auto-regressive terms, and the dimension p refers to the number of preceding time-steps to include in the auto-regressive parameter estimation. The parameters Θ refer to the moving-average terms, whose dimension, q, corresponds to the number of time steps included in the moving-average parameter estimation. Hence, the matrices to be estimated are $\Phi_1,\ldots,\Phi_p$ and $\Theta_1,\ldots,\Theta_q$, and $\boldsymbol{X}_t = (X_{1,t},\ldots,X_{k,t})^t$ denotes k-dimensional vector.

A VARMA model can be refined to include dependency among observations from neighboring locations (see Giacomini and Granger (2001) and references therein). Suppose there are N spatial locations, we introduce a spatial-correlation matrix $\Psi = [\psi_{k,j}] \in R^{N\times N}$. For each k, where $\sum_j \psi_{k,j} = 1$ and $\psi_{k,j}$ is nonzero only if location j is a neighbor of location k. The MSTAR-MA(p,q) model can hence be written as:

$$X_{k,t} - \sum_{d=1}^{p}\sum_{j=1}^{N}\phi_d\psi_{k,j}B^d X_{j,t} = a_{k,t} + \sum_{d=1}^{q}\sum_{j=1}^{N}\theta_d\psi_{k,j}B^d a_{j,t} \tag{3}$$

where $cov(a_{k,t}, a_{j,t'}) = \sigma^2 I(k=j)I(t=t')$. In a matrix representation where $\boldsymbol{X}_t = (X_{1,t},\cdots,X_{N,t}) \in R^N$ and similarly for $\boldsymbol{a}_t$, we have

$$\boldsymbol{X}_t - \sum_{k=1}^{p}\phi_k \Psi B^k \boldsymbol{X}_t = \boldsymbol{a}_t + \sum_{j=1}^{q}\theta_j \Psi B^j \boldsymbol{a}_t \tag{4}$$

The parameters to be estimated include $\phi_1,\ldots,\phi_p,\theta_1,\ldots,\theta_q,\Psi,\sigma^2$. There are therefore, in general, $T(p+q)N^2$ parameters to estimate in this formulation, where the dimension T corresponds to the number of time periods to be calibrated, each with its own set of transient parameters.

2.2.2. Leveraging road network characteristics to reduce the number of parameters

The basic transient model described in the previous section accounts for both spatial and temporal interactions, but does not respond to needs for parsimony in the model definition. To respond to that requirement, we make use of a decomposition of time into intervals, or templates, $r = 1\ldots R$, that permit combining time periods into like sets. An example of a decomposition is peak versus off-peak times of the day/week.

A shortcoming of all existing work in the area which takes into account neighboring links' effect on each link's traffic prediction is that the spatial correlations make use either of a fixed number of neighbors or, in the most general cases (e.g. Kamarianakis et al., 2005), are a function of distance. While distance may be suitable for modeling spatial correlations in some applications, it is clear that in transportation networks, the impact of one link's traffic characteristics on another link depends primarily upon the speed at which the traffic is traveling.

Hence, it is important to take into account the speed in the definition of the spatial-correlation matrix. Unfortunately, though, speed is what is being predicted by the model, so it is not possible to directly incorporate a functional dependence into the matrix.

Our approach, therefore, is to make use of the data history to induce not only a set of mean values for the speed and volume, but in parallel a set of spatial matrices. In other words, each reference period, $i = 1\ldots I$, has associated with it a spatial-correlation matrix which corresponds best, on average, to the relevant neighboring links during the period.

Let $S^{ri} \in \{0,1\}^{N\times N}$ be the i^{th} spatial correlation matrix, for template r. Then, the values of each S^{ri} reflect the links reachable in i time-steps in average conditions as reflected by template r. Hence, if a model includes, for each template r, two S matrices, S^{r1} and S^{r2}, then one would expect that the principal time-lag components in the estimation of the parameters, Φ, will be $t-1$ and $t-2$. Note however, that the definition of the spatial-correlation matrices are done once and are not estimated. Note also that it is quite reasonable to use the same decomposition of time for both R and I. One example of two possible values for R is peak versus off-peak.

The resulting parsimonious transient model is thus defined as

$$X_t - \sum_{l=1}^{p}\sum_{i=1}^{I}\left[\Phi_{lir}\otimes S^{ri}\right]X_{t-l,r} = a_t + \sum_{j=1}^{q}\sum_{i=1}^{I}\left[\Theta_{jir}\otimes S^{ri}\right]a_{t-j,r} \tag{5}$$

where I represents the number of spatial-correlation matrices that are computed in advance, $\otimes$ represents the Hadamard product of matrices, ie. entry-wise product of two matrices.

The number of parameters to estimate is therefore bounded by $IR(p+q)\gamma N$. The number I will typically be quite small, for example between 2 and 10. The parameter γ represents the maximum number of neighbors included for each link. Since I and R are fixed in advanced, the estimation problem scales well with increasing network size.

2.2.3. Fine structure of the model matrices

The adjacency matrix $S^{ri} \in \{0,1\}^{N\times N}$ in Eq. (5) considers the spatial correlations, which significantly reduces the number of parameters. Moreover, considerations of serial correlations could lead to further reduction of the number of parameters. The main vehicle is the Kronecker Indices of VARMA, see (Tsay, 1991). A brief illustration is as follows. For a 2-dimensional ARMA model, suppose its Kronecker Indices are identified as $K_1 = 2$, $K_2 = 1$, then it can be rewritten in the following Echelon form:

$$\Phi_0\begin{pmatrix}X_{1,t}\\X_{2,t}\end{pmatrix} - \Phi_1\begin{pmatrix}X_{1,t-1}\\X_{2,t-1}\end{pmatrix} - \Phi_2\begin{pmatrix}X_{1,t-2}\\X_{2,t-2}\end{pmatrix} = \Phi_0\begin{pmatrix}a_{1,t}\\a_{2,t}\end{pmatrix} + \Theta_1\begin{pmatrix}a_{1,t-1}\\a_{2,t-1}\end{pmatrix} + \Theta_2\begin{pmatrix}a_{1,t-2}\\a_{2,t-2}\end{pmatrix} \tag{6}$$

The coefficient matrices have special form:

$$\Phi_0 = \begin{pmatrix}1 & 0\\ * & 1\end{pmatrix}, \Phi_1 = \begin{pmatrix}* & 0\\ * & *\end{pmatrix}, \Phi_2 = \begin{pmatrix}* & *\\ 0 & 0\end{pmatrix}, \Theta_1 = \begin{pmatrix}* & *\\ * & *\end{pmatrix}, \Theta_2 = \begin{pmatrix}* & *\\ 0 & 0\end{pmatrix}$$

where * represents unknown parameters to be estimated. The identified zero elements in Φ_1, Φ_2, Θ_2 do not need to be estimated through an optimization technique, which simplifies the model estimation task. This fine structure of model matrices also benefits the prediction task in practical application. Since these zero entries effectively eliminate redundant dependency on past traffic information at certain links and time lags, the prediction step is therefore less susceptible to missing data in

real-time data stream. So the prediction step is more likely to produce stable output. Very often the variance of volume and speed at a given link varies during different periods of a day, such heteroskedasticity requires a more robust statistical procedure to identify the Kronecker Indices, see (Min and Tsay, 2005) for more details.

2.2.4. Example

Consider the following example road network. Links 1–10 are consecutive highway links, link 11 is a slip road (i.e., an on-ramp), and the remaining links are arterials. In this example, we let $r = 2$, so that we consider two distinct regimes, for example peak, or congested, and off-peak, or free-flow. In practice, r would be larger to account for time-of-day patterns, day-or-week behavior, and most likely holiday versus non-holiday periods. Furthermore, we set $i = 2$ here as well, for simplicity. Fig. 1

Then, suppose that the following average speeds have been obtained from the historical data for the two road types and the two distinct regimes. The first table represents values for highway links, and the second table represents values for arterials and for on- and off-ramps. In addition to these numbers of links which represent the downstream links contributing to a given link's traffic flow at time t, we shall suppose here that historical data has confirmed that in congested templates, two downstream highway links are included and one downstream arterial link, and none during free-flow conditions, due to the queueing effect of traffic flow and back propagation of the effect of congestion. This is not a statement of direct causality, but rather of correlation. Indeed, it is the case, in this example, that the flow, or speed, on link 11 is *correlated* with that of link 2, even though there is no direct flow of traffic from 11 to 2 or vice-versa. Correlation means that information about the state of link 2 has predictive power for that of link 11 and vice-versa. Since 11 represents a slip road for access onto the highway link, the correlation is not surprising.

Highway Free Flow Template, $r = 1$	*Raw data*	*Data on 5 mn-spaced intervals*
speed	120 km/h	10 km/5 mn interval
average link length	2 km	5 links traversed /5 mn interval
Highway Congested Template, $r = 2$	*Raw data*	*Data on 5 mn-spaced intervals*
average speed	72 km/h	6 km/5 mn interval
average link length	2 km	3 links traversed /5 mn interval
Arterials/Ramp Free Flow Template,$r = 1$	*Raw data*	*Data on 5 mn-spaced intervals*
speed	48 km/h	4 km/5 mn interval
average link length	1 km	4 links traversed /5 mn interval
Arterials/Ramp Congested Template, $r = 2$	*Raw data*	*Data on 5 mn-spaced intervals*
average speed	24 km/h	2 km/5 mn interval
average link length	1 km	2 links traversed /5 mn interval

In this example the matrices S^{r1}, for $r = 1,2$ are as follows. Subsequent matrices, S^{r2}, for example, contain 1 when the link can be reached by each link in two time-steps, given the granularity of the historical data.

Each row of an S matrix represents a link in the network. Note that the matrix is indeed square, since the columns also represent the link numbers. hence, the binary coefficients indicate, for each row (link) which other links have an impact on that link for the template in question, where the template is suggested by the parameters r and i.

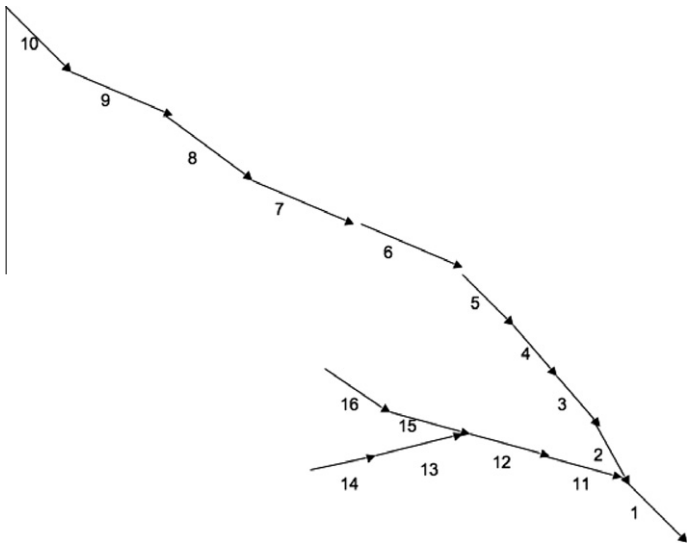

Fig. 1. Sample network with 10 highway links, five arterial links, and one on-ramp.

In this example, data is separated by 5 mn intervals, hence the second set of matrices contains 1 where the link is reachable in 10 min, and using some rule for upstream links as before. Note that dashes in the matrix implies that no data is available to calibrate the link's traffic characteristics corresponding to the row in question.

$$S^{11} = \begin{bmatrix}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\
2 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
3 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
4 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
5 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
6 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
7 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
10 & – & – & – & – & – & – & – & – & – & – & – & – & – & – & – & – \\
11 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\
12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\
13 & – & – & – & – & – & – & – & – & – & – & – & – & – & – & – & – \\
14 & – & – & – & – & – & – & – & – & – & – & – & – & – & – & – & – \\
15 & – & – & – & – & – & – & – & – & – & – & – & – & – & – & – & – \\
16 & – & – & – & – & – & – & – & – & – & – & – & – & – & – & – & –
\end{bmatrix}$$

$$S^{21} = \begin{bmatrix}
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 \\
1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
2 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
3 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
4 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
5 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
6 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
7 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
8 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
9 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
10 & – & – & – & – & – & – & – & – & – & – & – & – & – & – & – & – \\
11 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\
12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\
13 & – & – & – & – & – & – & – & – & – & – & – & – & – & – & – & – \\
14 & – & – & – & – & – & – & – & – & – & – & – & – & – & – & – & – \\
15 & – & – & – & – & – & – & – & – & – & – & – & – & – & – & – & – \\
16 & – & – & – & – & – & – & – & – & – & – & – & – & – & – & – & –
\end{bmatrix}$$

2.3. Predicting multiple time points into the future

One of the goals of our model is to predict speed and volume not only one time-step into the future, but several; in particular, the desired look-ahead prediction interval is from 5 min to 1 h. Since we make use of data arriving with a frequency of 5-min, our method is designed to be able to provide accurate forecasts up to 12 time points into the future.

Doing so is straightforward with the model we developed, by substituting the $t + 1$'st prediction as if it were an observed value and iterating again to predict speed and volume at $t + 2$. The procedure is repeated as such up to $t + 12$.

3. Test results and analysis

The proposed traffic prediction algorithm is implemented and tested against the actual traffic volume/speed over a medium size road network from the business district in an urban area on real-time basis. The extracted road network consists of

502 Links. The real-time traffic status are collected by loop detectors and summarized into 5-min volume and speed over each link. Data collected from some taxis are also included for inferring traffic status. Link length varies from 20 to 200 m.

All road types are included in the network and in the predictions. The expressway links are coded as Category A, Major Arterials as Category B, Standard Arterials are Category C, Minor Arterials are Category D, Small Local Roads are Category E, and On/Off-ramps are the Slip-Roads. The breakdown by road type in the test network is: 149 Category A, 246 Category B, 29 Category C, 38 Category D, 22 Category E and 18 Slip-Road.

The forecast up to 1 h ahead is issued every 5 min using the most recent actual traffic data. The forecasting accuracy for volume is measured by:

$$\text{Accuracy} = 1 - \frac{1}{T}\sum_{i=1}^{T}\left|\frac{forecast.vol(i) - vol(i)}{vol(i)}\right| \tag{7}$$

where T is the number of combinations of links and time points in the test dataset, and similarly for speed forecasting accuracy.

3.1. Numerical results of model fitting

Our MSTARMA(6,0) model of Eq. (3) was fitted to the de-trended speed or volume, see Eq. (1). We group 1 week into three classes and for each class we introduce two modes: daytime, which we refer to as "peak" and is defined as (7:00AM 8:00PM) and nighttime, called "off-peak" hereafter. The three day-of-week classes are: (Monday and Friday, Tuesday, Wednesday, and Thursday, Saturday and Sunday. Therefore we adopted six separate templates in total. The training set consists of recorded data from the past eight consecutive weeks. It was fitted through standard maximum likelihood estimation. Rather than showing full coefficient matrices for all 502 links with its corresponding huge dimension, we show the estimated model parameters for one link, link id 103088240, using 5-min average speed data, and corresponding to the template of Peak of

Table 1
Model parameters for link 103088240 corresponding to the peak template for the period of Monday and Friday, 7:00AM till 8:00PM.

Link-ID	Lag-1	Lag-2	Lag-3	Lag-4	Lag-5	Lag-6
103057453	0.01	−0.00	0.05	0.04	0.08	0.15
103070708	0.22	0.25	0.13	0.23	0.19	0.14
103088240	0.09	0.02	−0.07	−0.00	0.05	0.02

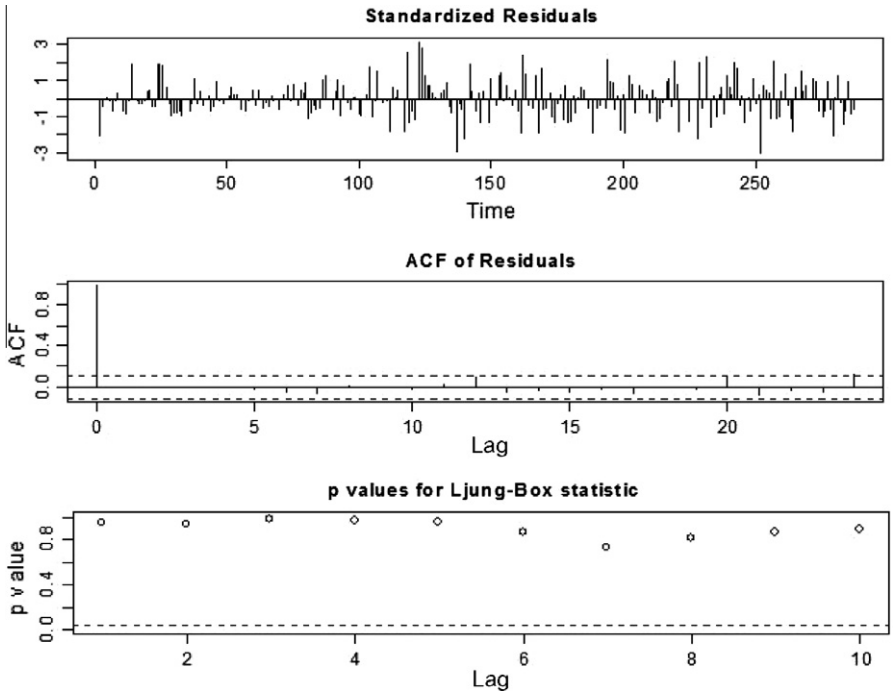

Fig. 2. Test residuals to check if there is any pattern. The test results confirms the residuals being white noise.

Tuesday, Wednesday and Thursday. It has two neighboring links out of the extracted road network (link id 103057453 and 103070708). The parameters for the selected link are reported in Table 1, broken down by time lag and neighboring link id (including the selected link id itself).

Quite interestingly, this table shows that the past volume or speed on neighboring links have strong impact on link 103088240. Particularly, the nearby neighboring link 103070708 has more time-lag impact (larger coefficients) than the link itself (compare the 2nd and 3rd row). Clearly it suggests that a univariate time series model for link 103088240 would be inadequate. Fig. 2 confirms the fitted model as being adequate, ie. no obvious pattern in the residuals.

3.2. Prediction performance

We report the prediction accuracy obtained in our experimental tests, ie. the prediction results were collected from a system implementation of the proposed model deployed as a pilot system in an urban city's traffic command center. Tables 2 and 4 summarize the average accuracy respectively for speed and volume, grouped by road category, from 7AM to 8PM on

Table 2
Average forecasting accuracy for volume from 7AM to 8PM on April 11 and 12, 2007.

Road category	Forecasting horizon					
	5 min	10 min	15 min	30 min	45 min	60 min
CATA	0.891	0.883	0.882	0.878	0.873	0.87
CATB	0.893	0.89	0.89	0.887	0.887	0.886
CATC	0.888	0.883	0.882	0.882	0.882	0.881
CATD	0.843	0.841	0.842	0.841	0.841	0.84
CATE	0.838	0.834	0.835	0.833	0.834	0.833
SLIP-ROAD	0.868	0.858	0.847	0.851	0.849	0.847

Table 3
Average forecasting accuracy for volume from 7AM to 8PM on April 13, 2007 (Friday).

Road category	Forecasting horizon					
	5 min	10 min	15 min	30 min	45 min	60 min
CATA	0.895	0.890	0.886	0.877	0.871	0.869
CATB	0.889	0.885	0.883	0.877	0.874	0.873
CATC	0.881	0.878	0.873	0.868	0.864	0.863
CATD	0.863	0.857	0.855	0.851	0.852	0.849
CATE	0.841	0.836	0.831	0.828	0.823	0.822
SLIP-ROAD	0.873	0.866	0.865	0.851	0.850	0.844

Table 4
Average forecasting accuracy for speed from 7AM to 8PM on April 11 and 12, 2007.

Road category	Forecasting horizon					
	5 min	10 min	15 min	30 min	45 min	60 min
CATA	0.95	0.943	0.94	0.93	0.923	0.92
CATB	0.873	0.874	0.873	0.874	0.867	0.874
CATC	0.875	0.875	0.875	0.876	0.867	0.875
CATD	0.851	0.852	0.853	0.852	0.834	0.855
CATE	0.828	0.828	0.83	0.832	0.791	0.83
SLIP-ROAD	0.922	0.915	0.921	0.912	0.907	0.911

Table 5
Average forecasting accuracy for speed from 7AM to 8PM on April 13, 2007 (Friday).

Road category	Forecasting horizon					
	5 min	10 min	15 min	30 min	45 min	60 min
CATA	0.946	0.939	0.938	0.929	0.922	0.917
CATB	0.857	0.859	0.857	0.850	0.843	0.842
CATC	0.839	0.840	0.837	0.831	0.829	0.828
CATD	0.830	0.830	0.828	0.823	0.822	0.823
CATE	0.814	0.815	0.812	0.807	0.801	0.801
SLIP-ROAD	0.921	0.916	0.913	0.903	0.897	0.896

April 11 and 12, 2007. Tables 3 and 5 report the same accuracy for the single day of April 13, 2007. April 11 and 12 of 2007 are Wednesday and Thursday and belong to the same day-of-week class, whereas April 13 of 2007 (Friday) belongs to another day-of-week class. There is little difference between Tables 2 and 3, similarly for Tables 4 and 5. This confirms stable performance for the whole period of April 11–13, 2007. Overall we see the accuracy for category D and E is less impressive compared to other categories. In fact, the input volumes and speeds on links of category D and E often have very high volatility over time due to various road factors, such as traffic light signals. Such wild behavior makes prediction a difficult

Table 6
Average forecasting accuracy for aggregated 15-min volume data (averaged over three time points from the 5-min volume data, for each link) from 7AM to 8PM on April 11 and 12, 2007.

Road category	Accuracy	Median	Standard deviation
CATA	0.922	0.929	0.024
CATB	0.933	0.938	0.021
CATC	0.928	0.931	0.024
CATD	0.903	0.912	0.03
CATE	0.878	0.874	0.029
SLIP-ROAD	0.905	0.917	0.044

Table 7
Average forecasting accuracy for speed from 7AM to 8PM of the week July 19–25, 2009.

Road category	Forecasting horizon					
	5 min	10 min	15 min	30 min	45 min	60 min
CATA	0.941	0.939	0.931	0.918	0.917	0.906
CATB	0.915	0.887	0.875	0.874	0.874	0.873
CATC	0.904	0.876	0.864	0.864	0.864	0.864
CATD	0.881	0.846	0.834	0.833	0.833	0.833
CATE	0.865	0.831	0.82	0.82	0.821	0.821
SLIP-ROAD	0.930	0.919	0.908	0.891	0.869	0.821

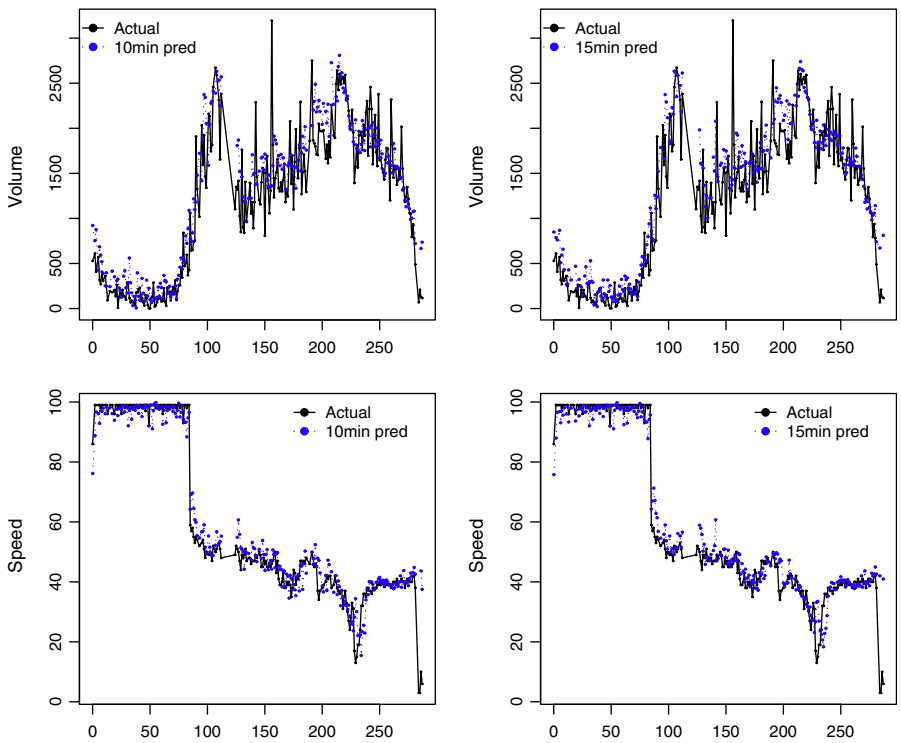

Fig. 3. Time series plot of volume and speed forecasts for one location on April 12, 2007. The *x*-axis represents the 288 5-min time intervals over one 24-h day. The *y*-axis represents hourly link volumes and 5-min average speeds, all recorded at 5-min intervals.

mission. In order to show that the good performance achieved in the field test in April, 2007 is not one-time event, we report in Table 7 the average speed forecasting accuracy for the week of July 19–25, 2009, which shows similar accuracy to the outcome of field test in 2007. It has been widely reported in the literature that simple historical average is inadequate for prediction of future traffic status. We have similar observations in our experimental tests, where the forecasting accuracy using simple historical averaging leads to on average an accuracy of about 10% lower than our results.

Fig. 3 has a time series plot of 10 min-ahead and 15 min-ahead forecasts against the actual speed and volume of one link on April 12, 2007. The *x*-axis is time over one 24-h day. Recall that we make use of 5-min real-time traffic data. There are 288 5-min intervals over a 24-h day, so that 0 on the *x*-axis corresponds to midnight, and 287 corresponds to 11:55pm the same day.

Fig. 4 shows the performance of speed and volume prediction during different 1-h periods of a day, average across April 11–13 of 2007. Overall the accuracy maintains stable and a high level from 7AM till 8PM.

The accuracy tables obtained on the test scenarios together with the plot illustrate the success of the implemented algorithm. Note that all accuracies are expressed as percentage accuracies in decimal form in terms of absolute relative error form the actual value as provided by the input data.

While we were interested in providing a robust, efficient, and accurate method to run continuously on 5-min traffic data feeds, it is useful to assess the accuracy of our method on 15-min averaged data. Indeed, 15-min data is less volatile and therefore easier to predict with high accuracy. In the majority of the papers published in the literature, 15-min averaged data is used. Table 6 presents three quantities measuring the 15-min volume forecasting performance by road categories: mean accuracy, median accuracy and standard deviation. Clearly high accuracy is achieved consistently across all road categories.

Since the method must be run continuously and the estimation of the model parameters cannot inhibit the real-time running of the model, the computation time of the method is a critical consideration.

Table 8 below shows the computation time needed for the estimation of the model parameters and the evaluation of the model, once estimated, for a week-long time horizon using a laptop computer (2.13 GHZ CPU and 2GB RAM). The network size is approximately 500 links. Since there are 288 5-min time points in a day, that gives 2016 time periods in the 7-day week, or over 1 million link $\times$ time points in all for a week.

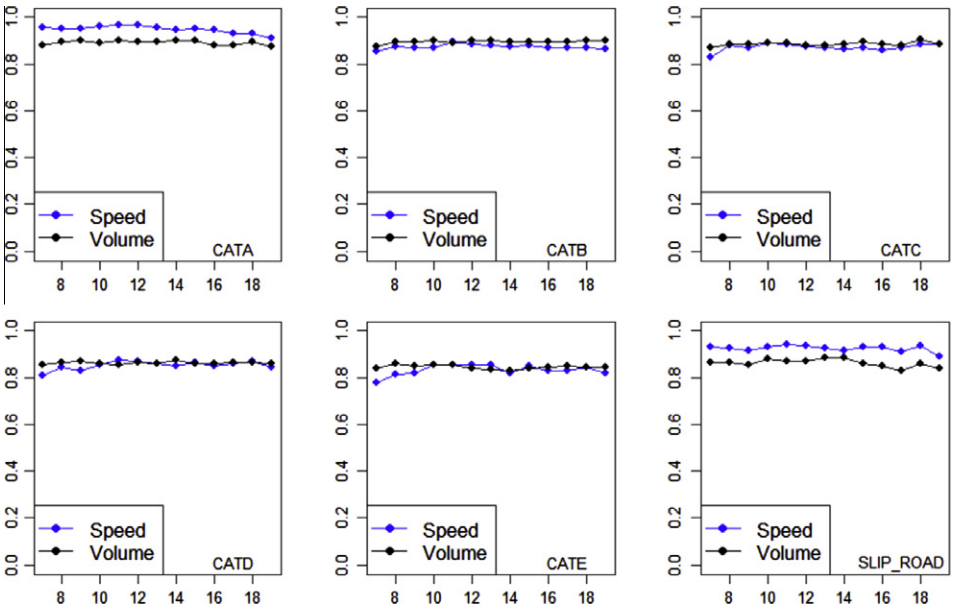

Fig. 4. Average forecasting accuracy during different one-h periods of April 11, 12 and 13, 2007, breakdown by road category and time-of-day (24-h clock).

Table 8
Computational time (in s) on a sparse network of 500 links with maximum neighbors 15 and maximum time lag 6.

Model parameter estimation	Single run in real-time using calibrated model, on a 500-link network
34	0.159

4. Conclusions and future work

The goal of this work was to develop a highly accurate and scalable method for traffic prediction at a fine granularity and over multiple time periods. That goal was clearly achieved by this effort. The accuracy exceeds that of other published work on 15-min data, and can achieve very good accuracy on the more volatile 5-min data. In addition, accuracy remains very good up to 12 5-min time periods into the future.

The method takes into account the spatial characteristics of a road network in a way that reflects not only the distance but also the average speed on the links. Because the method is designed to minimize the number of parameters needed to estimate, it remains computationally light and hence can be scaled to even large metropolitan areas.

Other aspects of this work involve incorporating weather, incident data, and roadwork, current or planned, into the forecasting model. The framework proposed in this paper can accommodate such factors in a coherent way. For instance, the factor of weather condition can be integrated in the stage of defining templates, more specifically, we can include weather condition together with day of the week and mode (peak, off-peak) to produce more templates. For incident or roadwork, if the network topology is changed (such as all-lane closure at certain links), then we may fit a new model with the new spatial adjacency matrix. Another interesting aspect is to create a simultaneous prediction model for volume and speed jointly. However, there are some practical factors that make this approach less attractive than it appears to be. The computer program processing signal from detectors could slow down if there is intense arrival of signal, which may occur in the mode of heavy volume and high speed. Such slowdown leads to longer data fusion processing time and therefore the volume it has processed may be incomplete when the reporting time is due. On the other hand, such incompleteness has little impact on reported speed due to large sample size of the processed signal (law of large numbers). The real-time volume is censored, therefore a simultaneous forecasting model for both speed and volume will be biased in this scenario. Another fact is often either speed or volume data is missing in the real-time stream on some links, which also makes separate model for speed and volume a more viable option from practical consideration.

References

A–Deek, H., Ishak, S., Wang, M., 2001.) A new short-term traffic prediction and incident detection system on I-4, vol. I. Final Research Report, Transportation Systems Institute (TSI), Department of Civil and Environmental Engineering, University of Central Florida.

Chandra, S.R., Al-Deek, H., 2009. Predictions of freeway traffic speeds and volumes using vector autoregressive models. Journal of Intelligent Transportation Systems 13 (2), 53–72.

Smith, B.L., Williams, B.M., Oswalsd, R.K., 2002. Comparison of parametric and nonparametric models for traffic flow forecasting. Transportation Research Part C 10 (4), 303–321.

Giacomini, R., Granger, C.W.J., 2001. Aggregation of space-time processes. Manuscript, Department of Economics, University of California, San Diego.

Kamarianakis, Y., Kanas, A., Prastacos, P., 2005. Modeling traffic volatility dynamics in an urban network. In: Transportation Research Record. Journal of the Transportation Research Board.

Kamarianakis, Y., Prastacos, P., 2003. Forecasting traffic flow conditions in an urban network: comparison of multivariate and univariate approaches. In: Transportation Research Record. Journal of the Transportation Research Board, 1857, pp. 74–84.

van Lint, J.W.C., Hoogendoorn, S.P., van Zuylen, H.J., 2005. Accurate freeway travel time prediction with state-space neural networks under missing data. Transportation Research Part C: Emerging Technologies 13 (5–6), 347–369.

Min, W.L., Tsay, R., 2005. On canonical analysis of multivariate time series. Statistica Sinica 15, 303–323.

Tsay, R., 1991. Two canonical forms for vector ARMA processes. Statistica Sinica 1, 247–269.

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

Wang, Y., Papageorgiou, M., Messmer, A., 2007. Real-time freeway traffic state estimation based on extended Kalman filter: a case study. Transportation Science 41 (2), 167–181.

Zheng, W., Lee, D.-H., Shi, Q., 2006. Short-term freeway traffic flow prediction: Bayesian combined neural network approach. Journal of Transportation Engineering 132 (2), 114–121.