Road Traffic Prediction with Spatio-Temporal Correlations

Wanli Min* Laura Wynter* Yasuo Amemiya*

The spatio-temporal relationship is an essential aspect of road traffic prediction. The fundamental observation is that the traffic condition at a link is affected by the immediate past traffic conditions of some number of its neighboring links. A time lag function defines how traffic flows are related in the temporal dimension. In parallel, the spatial structure defines which neighboring links have an effect on the traffic characteristics of other links, as a function of road type, speed, etc. This paper presents a new method which provides a complete description of the most important spatio-temporal interactions in a road network while maintaining the estimatability of the model. It improves upon existing methods proposed in the area and provides high accuracy on both urban and expressway roads.

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 controlers 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 occured prior to her starting her journey. The latter is particularly true for medium-length trips, such as 10-minutes 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-minute 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. This paper seeks to fill precisely that gap.

Using an approach that bears some similarity to that which we propose, Smith et al [1] performed a comparison of a seasonal ARIMA model and a nearest neighbor technique to predict 15-minute 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

*IBM Watson Research Center, PO Box 218, Yorktown Heights, NY, 10598 {wanli, lwynter,yasuo}@us.ibm.com

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.

In addition, there are a number of references in which real-time traffic prediction is performed using neural networks, such as [5, 7, 9]. The reference [9] examines 15-minute traffic volumes during daytime hours on an expressway, and attempts to predict the volume for the next 15-minute 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-minute 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 [4] 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 [4]. Furthermore, the method proposed by the authors in [4] 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 one day. This does not, however, deal with inter-day fluctuations, which should violate the stationarity assumption and introduce non-negligeable 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 [8] 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 upstream 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-minute average values) and handle predictions of one time point into the future. However, much real-time data is provided on a 5-minute basis. Hence, a method should be abe to react on that finer (and more volatile) time scale. In additional, 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 weekdays, weekends and holidays, with a high degree of accuracy on 5-minute speed and volume readings. Since the model is to be run in real-time and continuously, it is to be expected that there will at times be missing data. Indeed, in a real-time system, the data quality is quite variable from one moment to the next and in particular during off-peak and nighttime hours. The method therefore needs to be robust to the often volatile and in some cases missing data and computationally light enough to be run continuously and produce output in real-time without interruption.

2.1 Notation and Basic Relations

Let i 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

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

where μ_{itr} 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_{itr} 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:

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

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 $\Phi_1, \cdots, \Phi_p$ and $\Theta_1, \cdots, \Theta_q$, and $\boldsymbol{X}_t = (X_{1,t}, \cdots, X_{k,t})^t$ denotes k-dimensional vector.

A VARMA model can be refined to include dependency among observations from neighboring locations (see Giacomini & Granger [2] and references therein). Suppose there are N spatial locations, we introduce a spatial-correlation matrix $\Psi = [\psi_{i,j}] \in R^{N \times N}$. For each i, where $\sum_j \psi_{i,j} = 1$ and $\psi_{i,j}$ is nonzero only if location j is a neighbor of i according to appropriate definition. The MSTAR(p,q) model can hence be written as:

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

where $cov(a_{i,t}, a_{j,t'}) = \sigma^2 I(i=j) I(t=t')$. 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

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

The parameters to be estimated include $\phi_1, \cdots, \phi_p, \theta_1, \cdots, \theta_q, \Psi, \sigma^2$. 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. [3]), 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 aproach, 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 is peak versus off-peak.

The resulting parsimonious transient model is thus defined as

$$\sum_{l=1}^{p}\sum_{i=1}^{I}\Phi_{lir}S^{ri}X_{t-l,r} = a_t + \sum_{j=1}^{q}\sum_{i=1}^{I}\Theta_{jir}S^{ri}a_{t-j,r},$$

where I represents the number of spatial correlation matrices that are computed in advance.

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 Example

Consider the following example road network. Links 1–10 are consecutive highway links, link 11 is 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 congested and 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.

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, 2 upstream highway links are included and 1 upstream arterial link, and none during free flow conditions.

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

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

In this example the matrices S^{1r} and S^{1r}, for $r = 1, 2$ are as follows. Subsequent matrices, S^{2r}, for example, contains 1 when the link can be reached by each link in two time steps, given the granularity of the historical data. In this example, data is separated by 5mn intervals, hence the second set of matrices contains 1 where the link is reachable in 10 minutes, and using some rule for upstream links

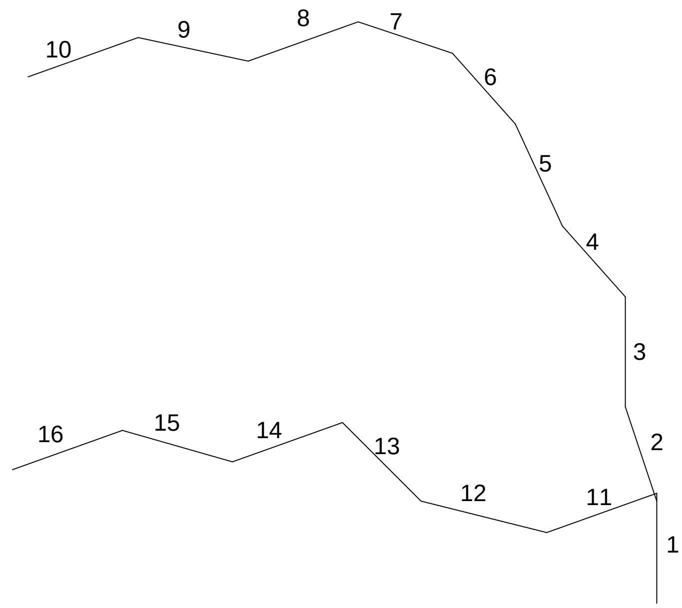

Figure 1: Sample nework with 10 highway links, 5 arterial links, and 1 off-ramp.

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.

$S^{11} =$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	0	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
3	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0
4	0	0	0	1	1	1	1	1	1	1	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	0
12	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
13	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
14	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
15	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
16	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

$S^{12} =$

	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	0	1	1	1	1	0	0	0	0	0	1	1	0	0	0	0
3	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0
4	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
6	0	0	0	0	0	1	1	1	1	0	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	1	1	1	1	0	0	0
12	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0
13	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
14	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
15	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
16	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

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 minutes to one hour. Since we make use of data arriving with a frequency of 5-minutes, 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 on real-time basis. The road network consists of 502 Links (149 Category A, 246 Category B, 29 Category C, 38 Category D, 22 Category E and 18 Slip-Road). The forecast up to one hour ahead is issued every 5 minutes using the most recent actual traffic data.

The forecasting accuracy for volume is measured by:

$$\text{Accuracy} = 1 - \sum_{i=1}^{N} \left| \frac{vol(i) - vol(i)}{vol(i)} \right| \tag{5}$$

and similarly for speed forecasting accuracy.

Table 1 and 2 summarize the average accuracy, grouped by road category, from 7AM to 8PM on April 11 and 12, 2007. Our model produces

Figure 2 has the time series plot of 10 minute-ahead, 15 minute-ahead forecast against the actual speed and volume of one link on April 12, 2007. The accuracy tables obtained on field test scenario together with the plot prove the indisputable success of the implemented algorithm.

Table 1: 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 2: 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

While we were interested in providing a robust, efficient, and accurate method to run continuously on 5-minute traffic data feeds, it is useful to assess the accuracy of our method on 15-minute averaged data. Indeed, 15-minute data is less volatile and therefore easier to predict with high accuracy. In the majority of the papers published in the literature, 15-minute averaged data is used. Table 3 presents

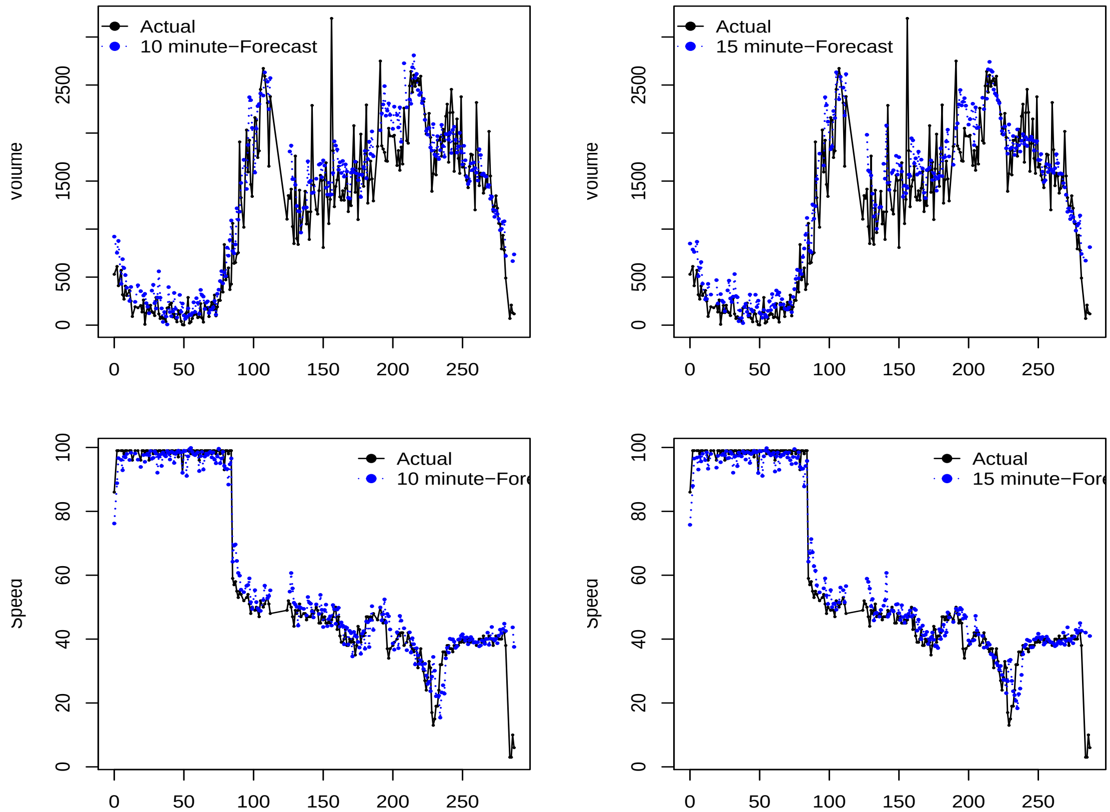

Figure 2: Time series plot of Volume and Speed forecast for one location on April 12, 2007.

three quantity measuring the 15-minute 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 model estimation cannot inhibit the real-time running of the model, the computation time of the method is a critical consideration.

The table 4 below shows the computation time needed for the estimation and the evaluation of the model 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-minute time points in a day, that gives 2016 time periods in the 7-day week, or over 1 million link×time points in all for a week.

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-minute data, and can achieve very good accuracy on the more volatile 5-minute data. In addition, accuracy remains very good up to 12 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

Table 3: Average Forecasting Accuracy for 15-minute Volume 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 4: Computational Time (seconds) on a sparse network of 500 links with maximum neighbors 15 and maximum time lag 6.

Estimation	Single real-time evaluation on a 500-link network
34	0.159

the number of parameters needed to estimate, it remains computationally light and hence can be scaled to even large metropolitan areas.

The next steps of this work are to incorporate weather, incident data, and roadwork, current or planned, into the forecasting model.

References

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

[2] Giacomini, R. and Granger, C.W.J. (2001). Aggregation of space-time processes. Manuscript, Department of Economics, University of California, San Diego.

[3] Kamarianakis, Y., Kanas, A., and Prastacos, P. (2005). Modeling Traffic Volatility Dynamics in an Urban Network. In *Transportation Research Record*, Journal of the Transportation Research Board.

[4] Kamarianakis, Y. and 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.

[5] 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.

[6] Pfeifer, P.E. and Deutsch, S.J. (1980) A three-stage iterative procedure for space-time modeling. Technometrics 22, 35-47.

[7] 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.

[8] Wang, Y., Papageorgiou, M., and Messmer, A. (2007) Real-Time Freeway Traffic State Estimation Based on Extended Kalman Filter: A Case Study. *Transportation Science*, 41(2), 167–181.

[9] Zheng, W., Lee, D.-H., and Shi, Q. (2006) Short-Term Freeway Traffic Flow Prediction: Bayesian Combined Neural Network Approach, *Journal of Transportation Engineering*, 132(2), 114–121.