Utilizing Real-World Transportation Data for Accurate Traffic Prediction

Bei Pan, Ugur Demiryurek, Cyrus Shahabi
Integrated Media System Center
University of Southern California
Los Angeles, United States
{*beipan,demiryur,shahabi*}*@usc.edu*

***Abstract*—For the first time, real-time high-fidelity spatiotemporal data on transportation networks of major cities have become available. This gold mine of data can be utilized to learn about traffic behavior at different times and locations, potentially resulting in major savings in time and fuel, the two important commodities of the 21st century. As a first step towards the utilization of this data, in this paper, we study the real-world data collected from Los Angeles County transportation network in order to incorporate the data's intrinsic behavior into a time-series mining technique to enhance its accuracy for traffic prediction. In particular, we utilized the spatiotemporal behaviors of rush hours and events to perform a more accurate prediction of both short-term and long-term average speed on road-segments, even in the presence of infrequent events (e.g., accidents). Our result shows that taking historical rush-hour behavior we can improve the accuracy of traditional predictors by up to 67% and 78% in short-term and long-term predictions, respectively. Moreover, we can incorporate the impact of an accident to improve the prediction accuracy by up to 91%.**

***Keywords*-traffic prediction, event impact analysis, time-series mining, transportation data**

I. Introduction

The two most important commodities of the 21st century are time and energy; traffic congestion wastes both. Several disciplines, such as in transportation science, civil engineering, policy planning, and operations research have studied the traffic congestion problem through mathematical models, simulation studies and field surveys. However, due to the recent sensor instrumentations of road networks in major cities as well as the vast availability of auxiliary commodity sensors form which traffic information can be derived (e.g., CCTV cameras, GPS devices), for the first time a large volume of real-time traffic data at very high spatial and temporal resolutions has become available. While this is a gold mine of data, the most popular utilization of this data is to simply visualize and utilize the *current* real-time traffic congestion on online maps, car navigation systems, sig-alerts, or mobile applications. However, the most useful application of this data is to predict the traffic ahead of you during the course of a commute. This predictive information can be either used by a driver directly to avoid potential gridlocks or consumed by a smart route-planning algorithm (e.g., [6]) to ensure a driver picks the best route *from the start*. According to a study by McKinsey Global Institute [1], using traffic information that avoids congestion can save consumers $600 billion annually by 2020.

In the past, several statistics, machine learning and data mining approaches have been applied to traffic data for prediction purposes, such as auto-regression [12], neural net [21] and smoothing [22] techniques. However, in this paper, we took a very pragmatic approach to evaluate and then enhance these techniques by intensely studying a very large-scale and high-resolution spatiotemporal transportation data from LA County road network. This dataset includes traffic flows recorded by under-pavement loop detectors as well as police reports on accidents and events. Our current system acquires these datasets in real time from various agencies such as Caltrans, City of Los Angeles Department of Transportation (LADOT), California Highway Patrol (CHP), Long Beach Transit (LBT), Foothill Transit (FHT) and LA Metro. In particular, for this paper, our main source includes approximately 8000 traffic loop-detectors located on the highways and arterial streets of Los Angeles County (covering 3420 miles, cumulatively) collecting several main traffic parameters such as occupancy, volume, and speed. [1]

Working with real-world data, we have identified certain characteristics of traffic data, such as temporal patterns of *rush hours* or the spatial impacts of *accidents*, which can be incorporated into a data-mining technique to make it much more accurate. For example, for generic time-series, the observations made in the immediate past are usually a good indication of the short-term future. However, for traffic time-series, this is not true at the edges of the rush hours. In that case, the historical observations (perhaps for that same day, time, and location) are better predictors of future. Hence, an auto-regression algorithm such as ARIMA [3], which by itself cannot capture sudden changes at the temporal boundaries of rush hours, can be enhanced by incorporating historical patterns.

While predicting short-term future has many applications, for example in fixing the errors of sig-alerts during rush-hours, it is not useful for smart path-planning where some-

[1]Even though this paper focuses on the sensor data collected from the loop detectors, our proposed techniques can be applied to other data collection approaches. For example, we can use the approaches proposed by [24] to aggregate the GPS data between regions, and consider the links between regions as *sensors* in our case.

times we need to know the traffic of a road-segment ahead of us by 30 minutes in advance. Again, historical data can improve long-term predictions because most probably the traffic behavior in 30 minutes at the desired location is similar to (say) yesterday's traffic at the same time and location. In this case, again ARIMA alone cannot be as effective since it only looks at immediate past and not the *right subset of* the historical patterns.

Unfortunately, even an enhanced ARIMA cannot predict accidents. However, if we know, from police event streams, that there is an accident (say, 30 minutes) ahead of us, we may be able to predict its delays and account for it. Again, historical data can be used to identify similar accidents, i.e., with similar severity, similar location and during the similar time, so that we can use their impact on average speed changes and backlog to predict the behavior of the accident in front of us. For example, our study shows that an accident that may happen between 4:00PM and 8:00PM on a particular segment of I-5 will cause 5.5 miles of average backlog ahead of the accident location. On the other hand, if the same accident happens between 8:00PM and midnight the backlog will be 2.5 miles.

The main challenge is how to properly incorporate all the knowledge from historical and real-time data into an appropriate time-series mining technique. This is exactly what we accomplished in this paper by enhancing ARIMA. Our experimental results with real-world LA data show that our enhanced ARIMA outperforms ARIMA by 78% when there is no unexpected events, and over 91% in the presence of events. In addition, we compared our enhanced approach with other competitor techniques for traffic prediction (e.g., [25] and [22]) and showed the superiority of our approach.

The main contributions of our work are:

- We analyze traditional prediction approaches based on real-world dataset, and discover their limitations at boundaries of rush hours, or in long term prediction. To overcome such limitations, we propose H-ARIMA approach which utilizes both historical traffic patterns and current traffic speed for prediction.
- We propose feature selection model to analyze the correlations between meta-attributes of traffic incidents (from event reports) and their impact areas (from traffic data). Later, we incorporate this model into our hybrid traffic prediction approach termed H-ARIMA+ to predict traffic in the presence of incidents.
- We evaluate our approaches with real-world traffic data, and event reports collected from transportation agencies, showing remarkable improvement in terms of prediction accuracy as compared with traditional traffic prediction approaches, especially at the boundaries of rush hours and at the beginning of unexpected traffic events, and for long term prediction.

The rest of this paper is organized as follows: Section II and III discuss related work and preliminaries, respectively. Section IV explains our enhanced ARIMA prediction approach. Section V describes our novel model to incorporate the impact of events in order to improve the prediction accuracy in the presence of events. Section VI reports our experiment results. Section VII, concludes the paper and discusses future plans.

II. Related Work

In this section, we review the related work on traffic prediction and event analysis techniques.

A. Traffic Prediction

The previous traffic prediction approaches can be grouped in two main categories: Simulation Models and Data Mining Techniques.

1) Simulation Models: The traffic prediction techniques developed in the first category use surveys and/or simulation models. In [5], Clark proposes a non-parametric regression model to predict traffic based on the observed traffic data. In [7] and [2], authors use microscopic models upon trajectories of individual vehicles to simulate overall traffic data and further conduct prediction. In [24], Yuan at el. estimate the traffic flow of a road segment by analyzing taxi trajectories. The major limitation of such studies is that they rely on sporadic observations and are often restricted to synthetic or simplified data for simulations.

2) Data Mining Techniques: The increase in the availability of real-time traffic allowed researchers to develop and apply data mining techniques to forecast traffic based on the real-world datasets. Since early 1980s, univariate time series models, mainly Box-Jenkins Auto-Regressive Integrated Moving Average (ARIMA) [3] and Holt-Winters Exponential Smoothing (ES) models [15], [22], have been widely used in traffic prediction. In the last decade, Neural Network (NNet) models also has been extensively used in forecasting of various traffic parameters, including speed [23], [10], travel time[21], and traffic flow [19], [17]. Nowadays, ARIMA, ES and NNet models are used as benchmarking methods for short-term traffic prediction [17], [16]. However, these approaches consider traffic flow as a simple time-series data and ignore phenomenons that particularly happen to traffic data. For example, for generic time-series, the observations made in the immediate past are usually a good indication of the short-term future. However, for traffic time-series, this is not true at the edges of the rush hours, due to sudden speed changes.

B. Traffic Event Analysis

The effect of events on traffic prediction has also been studied in the fields of data mining and transportation engineering. The majority of these studies focused on real-time event/outlier detection using probabilistic or rule-based approaches (e.g., [14], [9], [13]). There are also several studies that mainly concern the cause of the events, aiming

at how to design the network or re-direct the traffic flows to avoid the delay of events (e.g., [4], [20]). However, none of these studies incorporate events into traffic prediction techniques, and hence fail to provide realistic estimations in the presence of events. The focus of our study, on the other hand, is to integrate the impact of various events into forecasting models. The most relevant work to our study is the model proposed by Kwon and Varajya[11]. Their model utilizes a nearest-neighbor technique to detect cumulative delays and impact regions caused by traffic incidents. The impact regions are defined with fixed thresholds. However, the impact of events on traffic congestion varies based on space and time. For example, the impact region of an accident occuring during rush hour is usually more severe. Similarly, an accident at an inter-state street has a different impact region than that of a surface street. In this study, we consider such spatiotemporal characteristics of traffic events in training our models.

III. PRELIMINARIES

A. Problem Definition

Consider a set of road segments comprising n traffic sensors (e.g., loop detectors). We assume that at given time interval t (e.g., every minute), each sensor provides a traffic data reading, e.g., speed $v[t]$. We formulate the speed prediction problem as follows:

Definition 1: Given a set of observed speed readings V={ $v_i(j)$, i = 1,...,n; j= 1,...,t}, where i and j denotes a sensor and continuous time increments, respectively. The prediction problem is to find the set V={$v_i(j)$, j = t+1, t+2,...t+h} for each sensor i, where h denotes the **prediction horizon**. For example, h=1 refers to predicting the value of speed at $t+1$, where t represents the current time.

Definition 2: Short-term prediction and *long-term* prediction refer to prediction of speed when $h = 1$ and $h > 1$, respectively.

B. Baseline Approaches

In this subsection, we introduce two techniques that comprise the baseline of our prediction model, namely Auto-Regressive Integrated Moving Average (ARIMA) and Historical Average Model (HAM).

1) Auto-Regressive Integrated Moving Average (ARIMA): This model [3] is a generalization of autoregressive moving average model with an initial differencing step applied to remove the non-stationarity of the data. The model can be formulated as

$$Y_{t+1} = \sum_{i=1}^{p} \alpha_i Y_{t-i+1} + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i+1} + \varepsilon_{t+1} \quad (1)$$

where $\{Y_t\}$ refers to a time series data (e.g., the sequence of speed readings). In the autoregressive component of this model ($\sum_{i=1}^{p} \alpha_i Y_{t-i+1}$), a linear weighted combination of previous data is calculated, where p refers to the order of this model and α_i refers to the weight of $(t-i+1)$-th reading. In the second part ($\sum_{i=1}^{q} \beta_i \varepsilon_{t-i+1}$), the sum of weighted noise from the moving average model is calculated, where ε denotes the noise, q refers to its order and β_i represents the weight of $(t-i+1)$-th noise.

As shown in Equation (1), the predicted value mainly relies on the linear combination of the data that occurred before time t. This model can be directly used to predict the traffic speed data, when prediction horizon h=1. When h >1, we can iterate the prediction process h times by using the predicted value as the input to predict the next value.

2) Historical Average Model: Our rigorous analysis on real-world traffic sensor data reveals that there is a strong correlation (both temporally and spatially) present among the measurements of the single and multiple traffic sensor(s) on road networks. For example, the traffic condition of a particular road segment on Monday 8:30AM can be estimated based the average of last four sensor readings for the same road segment at 8:30AM in the past four Mondays. Therefore, we introduce Historical average model (HAM) that uses the average of previous readings for the same time and location to forecast the future data. We formulate HAM as follows:

$$v(t_{d,w} + h) = \frac{1}{|V(d,w)|} \sum_{s \in V(d,w)} v(s) \quad (2)$$

where $V(d,w)$ refers to the subset of past observations that happened at the same time d on the same day w. Specifically, d captures the daily effects (i.e., the traffic observations at the same time of the day are correlated), while w captures the weekly effects (i.e., the traffic observations at the same day of the week are correlated). For example, if the traffic data to be predicted is next Monday at 8:00AM, d refers to "8:00AM", and w = *Mon*. Thereby $V(d,w)$ refers to the set of traffic data happens on previous Mondays at 8:00AM. In fact, the selection of historical observations is also relevant with seasonal effects. For example, the historical observations on Mondays during *winter* is probably different with that on Mondays during *summer*. Here, we eliminate the seasonal effects by only using the data collected in one season. Also, as shown in the formula, the function to select past observations and calculating the average are indifferent to the value of the prediction horizon h.

C. Case Studies

One can use either ARIMA or HAM for traffic prediction in road networks. Here, we explain the limitations of both techniques based on our observations derived from real-world traffic datasets. Towards that end we present two case studies using different prediction horizons and temporal scales (i.e., rush hour boundaries).

1) Effect of Prediction Horizon (h): In the this case study, we would like to compare the prediction accuracy of ARIMA and HAM for different prediction horizons using real-world traffic data. (see Section VI for details of the real-world dataset and experimental setup). Note that the

aggregation level for this data set is 5 mins. Our intuition is that ARIMA relies on the very recent traffic data, which are usually a good indication of the near future. On the other hand, HAM uses the average of historical data for prediction, and hence HAM is more accurate in long-term prediction and its accuracy is independent of the prediction horizon. Our hypothesis can be summarized as follows:

Hypothesis 1: The prediction horizon has no noticeable effect on the prediction accuracy of HAM. However, as the prediction horizon increases, the prediction accuracy of ARIMA decreases.

The result of comparison using real data is presented in Figure 1, which measures the average mean absolute percentage error of prediction (y-axis) with respect to prediction horizon (x-axis). As shown in Figure 1, ARIMA yields better prediction than that of HAM when $h < 6$ (i.e., less than 30-min in advance prediction). However, as h increases to the values larger than 6, HAM starts to yield better prediction. This result not only verifies hypothesis 1, but also reveals that ARIMA is not ideal for long-term predictions (i.e., more than 30-min in advance prediction).

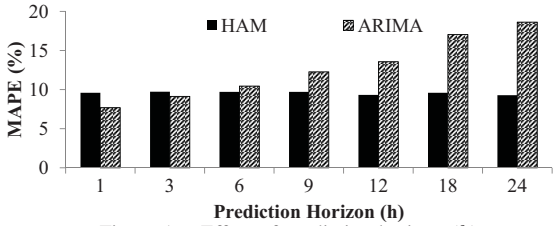


Figure 1. Effect of prediction horizon (h)

2) Effect of Rush Hour Boundaries: The intuition here is that the observations made in the immediate past are usually a good indication of the short-term future. Therefore ARIMA is excepted to yield accurate prediction in the short-term. However, the speed change at rush-hour boundaries is sudden and there is no indication (i.e., trend) of such change before it happens. In such cases, ARIMA cannot capture the speed changes at the very beginning, but adjust itself shortly after it takes the changed speed into account. On the other hand, since rush hours happen at almost same time of that particular day, HAM can predict the sudden speed changes at the boundary of rush hours. Our intuition can be summarized with the following hypothesis:

Hypothesis 2: HAM can efficiently predict the sudden speed changes at the boundaries (i.e., beginning and end) of rush hours. On the other hand, ARIMA has a delayed reaction on the boundaries.

In this case study, we fix the prediction horizon (i.e., h=6) and compare the prediction accuracy of both approaches over time using real-world traffic speed data. The experimental results are depicted in Figure 2, which represents the actual speed data and predicted values from two models for a specific sensor at different times of a particular weekday. As shown, in the morning rush hour around 6:50AM, HAM predicts the beginning of congestion with a very small error rate and ARIMA's prediction is shifted (with respect to actual speed) a few timestamps. Similarly, at the vanishing point of the rush hour congestions around 9:05AM, HAM still accurately predicts the after-congestion speed and ARIMA shifts a few timestamps. The results show that at the boundaries of rush hours, HAM yields higher prediction accuracy than that of ARIMA. Hence, the Hypothesis 2 is verified.

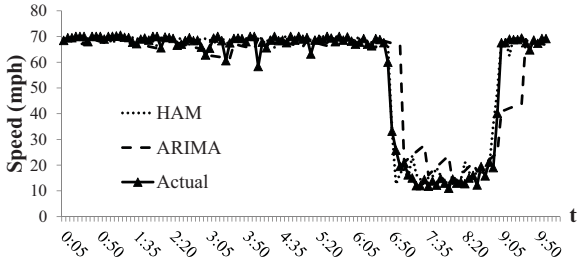


Figure 2. Effect of rush-hour boundaries

IV. Hybrid Forecasting Model

In this section, we propose a hybrid forecasting model named Historical ARIMA (H-ARIMA) that selects in real-time between ARIMA or HAM based on their accuracy. In particular, as the traffic data streams arrive we compare the accuracy of ARIMA and HAM, and select the one that yields low prediction error. As we discussed ARIMA relies on the very recent traffic data, and hence in some circumstances (i.e., in the long-term when $h \geq 6$ and at the boundaries of rush hours) its prediction accuracy degrades significantly. On the other hand, HAM uses past observations to predict future traffic conditions. While HAM yields better prediction for long-term, it is not ideal for short-term predictions. Therefore, the main idea behind our hybrid approach is to distinguish the circumstances when a specific approach is better. Towards that end we train a decision-tree model that selects between ARIMA and HAM to forecast the speed at individual time stamps. In this model, the decision parameter and threshold are denoted as λ and ϕ, respectively. For each time stamp t, we choose between ARIMA and HAM based on the trained value of λ_t. If $\lambda_t \leq \phi$, we choose ARIMA, otherwise, we choose HAM. The value of λ_t is calculated based on the rate of overall prediction error between HAM and ARIMA at t. The detailed approach is described in Algorithm 1, given the entire training dataset $\{v(j)\}$ (j=1...t), together with the value of d and w.

In Line 1 of Algorithm 1, we initialize dataset *S* with all the historical data observed on day w, at time d. For example, if w = *Mon* and d = *8:00AM*, the set of *S* refers to all the traffic speed readings on Mondays at 8:00AM within the training dataset. In Line 4-9, we utilize ARIMA and HAM to *predict* speed reading v_i in *S* and compute

Algorithm 1 Get $\lambda(\{v(j)\}$, d, w)

Output: λ

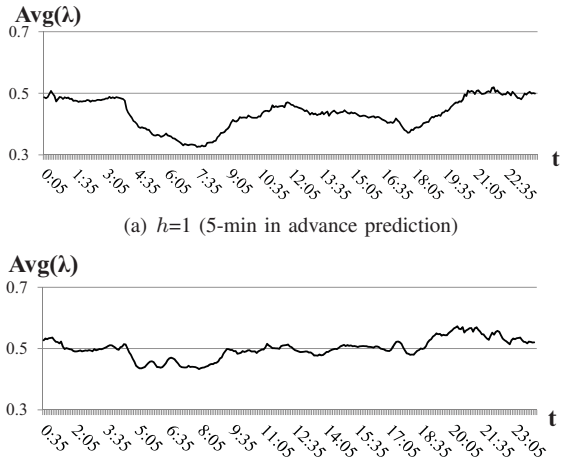
1: Let S = $\{V(\{\text{v(j)}\}, d, w)\}$
2: Let Err_{ARIMA} = 0; $Err_{\text{HAM}} = 0$
3: Initialize ARIMA model with training dataset $\{v(j)\}$
4: v_{HAM}= Average(V{d,w});
5: **for all** $v_i \in S$ **do**
6: v_{ARIMA} = ARIMA(i);
7: Err_{ARIMA} = Err_{ARIMA}+ RMSE(v_i, v_{ARIMA});
8: Err_{HAM} = Err_{HAM}+ RMSE(v_i, v_{HAM});
9: **end for**
10: λ = Err_{ARIMA} / (Err_{ARIMA}+Err_{HAM})
11: Return λ.

their prediction error. In Line 10, λ is calculated as the ratio of the prediction error from ARIMA versus the sum of prediction errors from two approaches. Based on the calculation strategy of λ in Algorithm 1, we observe that if λ <0.5, the total prediction error from ARIMA is less than that of HAM, which means ARIMA is better for this particular time stamp (i.e., time d on day w). Otherwise, HAM is better. Thereby, we set threshold ϕ as 0.5.

To further explain the robustness of H-ARIMA, we present the training results for λ in the following two main cases.

First, we study the effect of d on λ. Figure 3 shows the effect of d with respect to the average λ from all sensors for two different prediction horizons: h=1 and h=6. Here, the day parameter w is fixed as *Wed*. Figure 3(a) indicates that in short-term prediction (i.e., h=1), the ARIMA yields better performance, because most average λ values are less than 0.5. Figure 3(b) shows that when h=6, there are more time instances with $\lambda > 0.5$. This indicates that HAM starts to provide better prediction accuracy in the long term (Hypothesis 1). In addition, both charts in Figure 3 show that during the morning and afternoon rush hours (i.e., 6:00AM to 9:00AM, 4:00PM to 7:00PM), the accuracy of HAM is not as good as compared to non-rush hours, reflecting that the average λ declines during the rush-hour interval. One possible explanation is that during rush hours, the impact of the unexpected events (e.g., accident) is more significant than that of non-rush hours. Since the effects of traffic accidents are offset by averaging the entire history, HAM cannot capture such effects. We will address this problem in Section V.

Second, based on the Hypothesis 2, we plan to examine behaviour of λ at the boundaries of rush hours, thereby we focus on the values of λ for a particular sensor. In Figure 4(a), we plot individual λ value for a single sensor over all daily time stamps(d). To analyze the behavior of λ over time, the historical average speed sequence is also plotted in Figure 4(b). Here, the prediction horizon is fixed to h=1,

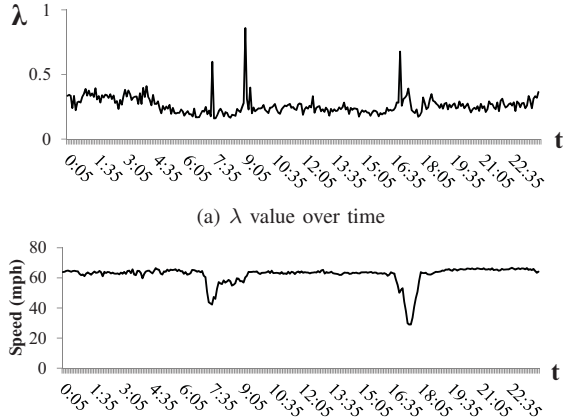


(a) h=1 (5-min in advance prediction)

(b) h=6 (30-min in advance prediction)

Figure 3. Effects of prediction horizons over average λ

and weekly parameter w = *Wed*.



(a) λ value over time

(b) Historical average over time

Figure 4. Effects of rush-hour boundaries over λ

In Figure 4(a), there are three time instances where $\lambda > 0.5$ (i.e., 6:35AM, 8:55AM and 4:35PM). As shown in Figure 4(b), those three time instances are exactly at the boundaries of rush hours. As indicated in Observation 2, at beginning and ending of the rush hours, HAM model outperforms ARIMA, even though the prediction horizon is only 1.

V. Event Impact Analysis

Traffic events include non-recurring incidents (e.g., accident, vehicle breakdown, and unscheduled road construction) which result in traffic congestion or disruption. In addition, we can consider social events such as a music concert at LA Live or Lakers basketball game at Staples Center. In this section, we study the effect of events on traffic congestions, especially in upstream direction. In particular, we incorporate event information in to H-ARIMA to enhance the prediction accuracy of our model. Towards this end, we exploit our historical event reports and the associated traffic speed nearby at the time of the events to model the correlation between event attributes and traffic congestion.

Note that even though our model is built offline by using the past data, we use it online for better traffic prediction. That is, in real-time using the current event reports as input, we match the event's attributes to find similar events happened in the past to predict speed delays and backlogs, caused by the current event.

As discussed in Section III, HAM can hardly react to unexpected traffic events as it eliminates the influence of events by averaging historical observations. ARIMA, due to its delayed reaction, is not an ideal method to use in the case of events which cause sudden changes in the time-series data. To illustrate the prediction accuracy of ARIMA and HAM in the presence of an event, consider Figure 5 that shows the speed prediction of both techniques for a traffic accident that happened on freeway CA-91 at 10:53AM Dec. 5th, 2011 with prediction horizon h =6. As shown, the prediction accuracy of both techniques are significantly low as compared with the actual speed.

Hence, we discuss our Event Impact Area (EIA) model that addresses traffic prediction problem in the presence of events.

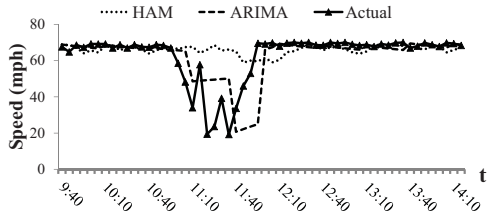


Figure 5. Impact of an accident on ARIMA and HAM

A. Event Impact Area Model

With our approach we assume that event data is an input to our algorithm and includes but not limited to the following meta-data: 1) event date, 2) event start-time, 3) event location (i.e., latitude, longitude), 4) event type (e.g., traffic collision, road construction), 5) type of vehicles involved if incident is an accident, 6) number of affected lanes. In addition, we introduce a parameter, namely impact post-mile, to represent the spatial upstream span of an event.

Definition 3: Impact post-mile is the distance between the location of an event and its last influenced sensor on the opposite direction of vehicle flow, as shown in Figure 6.

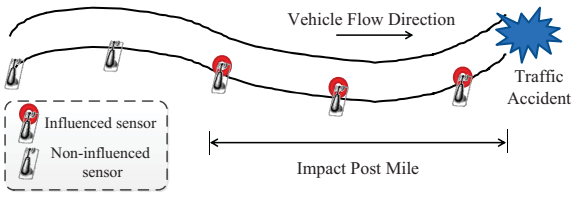


Figure 6. Definition of event impact post-mile

The influenced sensors are the sensors whose speed reading show an *anomalous* decline compared with the historical average speed[2].

[2]We detect the *anomalous* decline using the traffic event detection algorithm proposed in [13].

Table I
AVERAGE IMPACT POST-MILE ON EVENT META-ATTRIBUTES

(a) Traffic collision event, affected lanes = 0

Location	D	$S_{0,4}$	$S_{4,8}$	$S_{8,12}$	$S_{12,16}$	$S_{16,20}$	$S_{20,24}$
I-405	N	2.07	2.93	3.68	2.92	3.33	1.51
I-405	S	0.14	3.37	2.61	3.63	4.37	2.03
I-5	N	0.10	3.32	4.12	4.45	5.51	2.56
I-5	S	1.17	3.66	3.41	2.43	3.73	1.34

(b) Traffic collision event, affected lanes = 1

Location	D	$S_{0,4}$	$S_{4,8}$	$S_{8,12}$	$S_{12,16}$	$S_{16,20}$	$S_{20,24}$
I-405	N	N/A	N/A	4.74	3.57	3.52	0.46
I-405	S	N/A	N/A	N/A	N/A	4.78	1.75
I-5	N	N/A	N/A	2.02	N/A	6.11	N/A
I-5	S	0.10	N/A	N/A	N/A	N/A	N/A

(c) Road construction event, affected lanes = 1

Location	D	$S_{0,4}$	$S_{4,8}$	$S_{8,12}$	$S_{12,16}$	$S_{16,20}$	$S_{20,24}$
I-405	N	0.96	N/A	9.35	5.02	N/A	1.25
I-405	S	1.73	N/A	N/A	N/A	N/A	0.19
I-5	N	N/A	N/A	4.70	5.80	5.70	6.50
I-5	S	N/A	N/A	N/A	N/A	N/A	N/A

Based on our analysis of real-world data, we observe that impact post-mile varies across events with different attributes. Let us consider one of the attributes "*start time*" as an example. The impact post-mile of events that happen during day-time may be large compared with events happening at midnight, due to higher traffic flow during the day-time. Thereby, the key to investigate the correlation between event attributes and impact post-mile is to decide which attributes are correlated with impact post-mile. It is likely that some event attributes are irrelevant or redundant for inferring impact post-mile. In order to identify the most correlated subset, we first process the event attributes as normalized features and impact post-mile as numerical classes, and then apply the Correlation based Feature Selection (CFS) algorithm[8] on top of this normalized data to select correlated features. From the result obtained from this procedure, we observe that the following event attributes are the most relevant:{*Start time*, *Location*, *Direction*, *Type*, *Affected Lanes*}. We use the selected attributes to classify the impact post-mile values, and utilize the average value of each class to represent the impact of an event with corresponding attributes. Table I shows some selected classification results where the impact post-mile under different *Start-time* is aggregated into 4-hour interval denoted as $S_{\text{start-hour,end-hour}}$ and "N/A" means that there is no such event happening with the attributes specified in our experimental dataset[3]. The dataset used to train this model includes the events happened in weekdays, when rush-hour is considered as 6:00AM to 9:00AM and 4:00PM to 7:00PM.

From the results shown in Table I, we make the following observations.

- First, from Table I(a), we observe that for the events

[3]The number of affected lanes equals zero indicates that no lanes are blocked as the involved vehicles moved to the shoulder of the road after the accident.

happening during rush hours, the impact post-mile is larger than that of non-rush hours. This is expected because when an accident happens during rush hours on a high occupancy road, the impact of that event is more severe than on roads without traffic.

- Second, comparing Table I(a) and I(b), we infer that for the events happening at similar time, same location, the impact post-mile is generally larger when the number of affected lanes is more. Obviously, since the affected number of lanes reflects the number of lanes which are blocked by the events, the more lanes blocked, the slower the traffic flow. However, for accidents that occur at midnight, since the traffic is free-flow at that time, the higher number of affected lanes does not necessarily indicate longer impact post-mile.
- Third, in Table I(c), we observe that for the road construction events, if they happen at day time, especially at rush hours, their impact on traffic is severe, sometimes exceptionally larger than that of traffic collisions happening at the same time. On the other hand, if they happen at night, their impact is not that significant.

B. Event Impact Prediction

In addition to impact post-mile, the speed change (speed-impact) caused by events is also very important for traffic prediction. To estimate the speed-impact, we introduce two factors: *influenced speed decrease* (Δv) and *influenced time shift* (Δt). We estimate Δv based on the correlated attributes (similar to impact post-mile).

Definition 4: For sensor i, its influence speed decrease Δv_i for event e is defined as the average speed changes for all events that share the same correlated attributes (i.e., *Start-time*, *Location*, *Direction*, *Type* and *Affected Lanes*) with e, and affected sensor i in the past.

Once we find the influenced speed decrease, the next step is to determine the exact time stamps we need to apply the change on sensors. When an event occurs, the sensors located at different locations might be influenced at different time stamps. Therefore, we propose the concept of *influenced time shift* (Δt) to estimate the period of time that a sensor will be affected after an event.

Definition 5: For sensor i, its *influenced time shift* (Δt_i) for event e is defined as the distance between the sensor i and event e divided by the average traffic speed between them, which can be represented as follows:

$$\Delta t_i(e) = \frac{dist(i,e)}{avg(\{v_j\})} \quad \text{where} \quad p(i) \leq p(j) \leq p(e) \qquad (3)$$

where *p(i)* refers to the post-mile of sensor i. The set of $\{v_j\}$ refers to all the speed readings presented at the sensors located between sensor i and event e.

Below we summarize our procedure to predict traffic in case of events:

Table II
DATASET DESCRIPTION

Sensor Data	duration	Nov. 1st - Dec. 7th, 2011
	# of sensors	2028
	spatial span	3420 miles
	sensor sampling rate	1 reading per 30 secs
	temporal aggregation interval	5 mins
	spatial resolution	1 sensor
Event Data	# of events	3255
	# of event attributes	43

1) When an event e occurs at time t, all the relevant event features(i.e., {*Start-time*,*Location*, *Direction*, *Type*, *Affected Lanes*}) are incorporated in the EIA model to determine the impact post-mile of e.
2) Using the impact post-mile and the location of e, the set of all influenced sensors are identified as set $\{s_i\}$.
3) For each sensor s_i, during [t+$\Delta t_i(e)$, t+$\Delta t_i(e)$+h], the predicted value is calculated as $(v_i(t) - \Delta v_i)$, where h is the prediction horizon.
4) After time t+$\Delta t_i(e)$+h, ARIMA is used to predict the rest until the event e is cleared.

VI. PERFORMANCES EVALUATION

A. Experimental Setup

1) Traffic Dataset: In our research center, we maintain a very large-scale and high resolution (both spatial and temporal) traffic loop detector dataset collected from entire LA County highways and arterial streets. We also collect and store traffic event data from City of Los Angeles Department of Transportation and California Highway Patrol. The detailed description of this dataset is shown in Table II.

2) Baseline Approaches:

- *ARIMA*: We implement ARIMA [3] starting with stationary verification, followed by the iterations of 1 to 10 for Auto Regressive model and 1 to 10 for Moving Average model to reach the best combination under Bayesian information criteria [18]. We use the trained model for one-step ($h = 1$) forecasting. When $h > 1$ (i.e., long-term forecasting), we iterate the prediction procedure for h times by using predicted value as previously observed value.
- *ES*: We implement Exponential Smoothing(ES) method as a special case of ARIMA model, with the order auto-regressive model set to zero, and the order moving average model set to 2.
- *NNet*: We implement Neural Network (NNet) model as multilayer perceptron (MLP). The architecture of MLP is as follows: 10 neurons in the input layer, single hidden layer with 4 neurons and h output neuron, where h refers to the prediction horizon. For example, in one-step forecasting, there is 1 output neuron. The input neurons include $\{v(k), k = t-9, ..., t\}$, while the output neuron is $\{v(t+1)...v(t+h)\}$, where t represents the current time. Tangent sigmoid function and linear transfer function are used for activation function in the

hidden layer and output layer, respectively. This model is trained using back-propagation algorithm over the training dataset.

3) Fitness Measurements: We use mean absolute percent error (MAPE) and root mean square error (RMSE) to quantify the accuracy of traffic prediction.

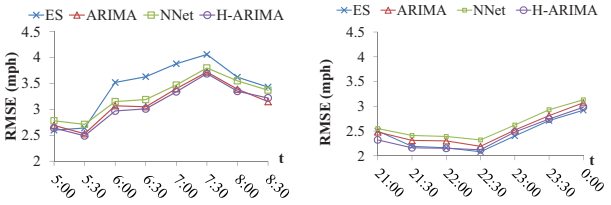
$$\text{MAPE} = \left(\frac{1}{N}\sum_{i=1}^{N}\frac{|y_i-\widehat{y}_i|}{y_i}\right)\times 100$$
$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\widehat{y}_i)^2} \quad (4)$$

where y_i and $\widehat{y}_i$ represent actual and predicted traffic speed respectively, and n represents the number of predictions.

B. Predictions Without Event Information

In this set of experiments, we use the traffic dataset collected from Nov. 1 to Nov. 30 as the training set. The dataset from Dec. 1 to Dec. 7 is used as testing set.

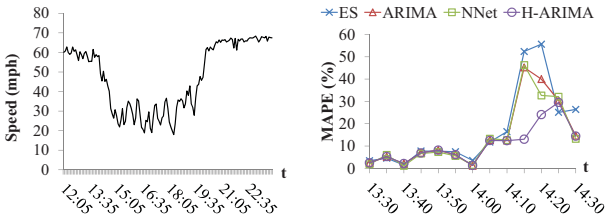
1) Short-term Prediction: In this experiment, we evaluate the short-term prediction (i.e., $h = 1$) accuracy of H-ARIMA with respect to baseline approaches. Figure 7 plots the average one-step prediction accuracy over all sensors on a weekday, 7(a) and 7(b) correspond to rush hour time interval and non-rush hour time interval, respectively. As shown, the accuracy of all prediction approaches during rush hour are lower than that of non-rush hours.



(a) Rush hour (b) Non-rush hour

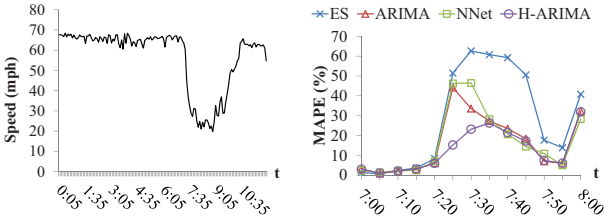
Figure 7. Overall RMSE (h=1)

Though H-ARIMA outperforms baseline approaches in general, it does not show clear advantages over them according to the aggregated results (over 2028 sensors). However, shown with the following experiment, H-ARIMA does have significantly better prediction accuracy than baseline approaches in the boundaries of rush hours. Figure 8 and 9 show the actual speed and MAPE of the prediction on two different road segments of I-5 and I-10. In 8(a), we observe that there is a sudden speed decrease around 14:00. Consequently, as shown in 8(b) at 14:15, we observe a significant increase in the prediction error of baseline approaches. This is because the baseline approaches cannot detect the sudden speed decrease in advance. On the other hand, H-ARIMA can estimate the beginning of congestion from historical pattern and yields better prediction by improving the baseline approaches up to 67.0% (at 14:15). Similarly, as shown in Figure 9(a) and Figure 9(b), the morning rush hour of I-10 starts around 7:00AM and H-ARIMA outperforms baseline approaches up to 61% (at 7:25AM). We note that this set of experiments focus on one-step forecasting where the baseline approaches can adjust themselves by utilizing the decreased speed, thereby their prediction accuracy recovers shortly.



(a) Actual speed (b) MAPE of the road

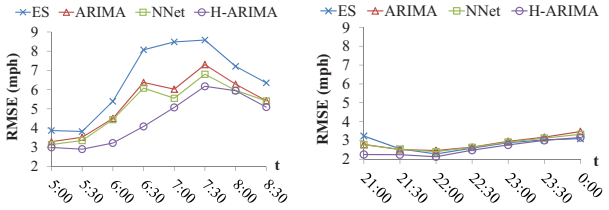
Figure 8. Case study on I-5 S. segment from Downtown



(a) Actual speed (b) MAPE of the road

Figure 9. Case study on I-10 W. segment to West-LA

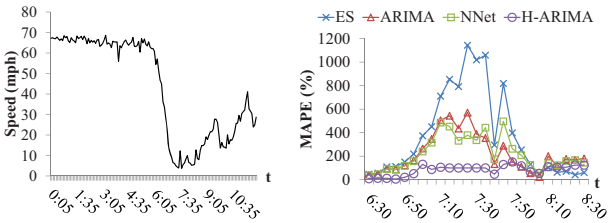
2) Long-term Prediction: In this set of experiments, we compare the prediction accuracy of H-ARIMA with baseline approaches for $h > 1$. Figure 10 plots the average six-step (i.e. $h = 6$) prediction accuracy over all sensors on a same weekday. Figure 10 shows that when prediction horizon increases, the prediction errors of baseline approaches increase, especially during rush hours (see Figure 10(a)). In Figure 10(a), we observe that H-ARIMA yields better prediction accuracy than that of baseline approaches. Similar to one-step prediction, in the next set of experiment we present the performance of H-ARIMA based on a road segment with rush hour congestion.



(a) Rush hour (b) Non-rush hour

Figure 10. Overall RMSE (h=6)

Figure 11 illustrates the actual speed and MAPE of the prediction on road segments of I-10. As shown in Figure 11(a), around 7:00AM, the speed decreases from 65 mph to 5 mph within a very short time. The baseline approaches can only sense this change with 30 minutes delay, and hence their MAPE is considerably high (see Figure 11(b)). On the other hand, H-ARIMA utilizes the historical congestion information to predict the traffic and hence its MAPE is fairly low as compared to baseline approaches. In particular, H-ARIMA improves the best baseline approach 78% (at 7:10AM).



(a) Actual speed (b) MAPE of the road

Figure 11. Case study on I-10 E. segment to Downtown

Table III
RELEVANT EVENT ATTRIBUTES

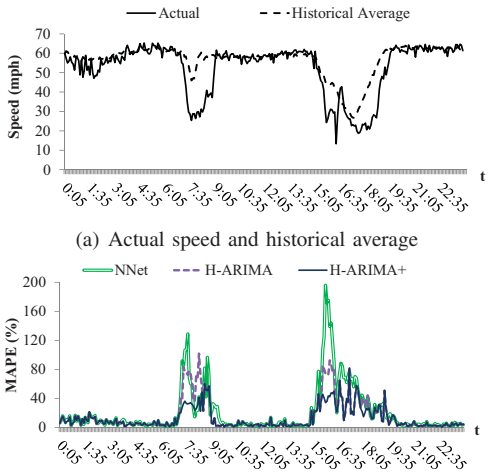
Event ID	Start time	No. of Affected Lanes	Dist(e,s)
350	06:31	0	0.58
2116	16:06	0	0.10
2621	18:26	0	0.11

C. Predictions with Event Information

In this set of experiments, we evaluate the prediction accuracy of our proposed approach in the case of events, dubbed H-ARIMA+ (see Section V). We compare H-ARIMA+ with H-ARIMA, and the best baseline approach in multi-step prediction (i.e., NNet). We set the prediction horizon of all approaches to 6, which indicates that our algorithm is set to predict speed information 30-minute in advance.

Figure 12 shows the result for a sample sensor located on east bound of CA-91 affected by three traffic collision events on Dec. 7, 2011. Figure 12(a) illustrates the actual speed on that day and the historical average (for that weekday) of the selected sensor. The historical average indicates that the rush hour intervals for this sensor are [7:00AM-8:00AM], and [3:00PM-7:00PM]. Figure 12(b) plots the prediction error for H-ARIMA+, H-ARIMA, NNet correspondingly. Table III shows the relevant attributes for each event, where *Dist(e,s)* refers to the distance between the sensor and corresponding event location. The number of affected lanes equals zero indicates that no lanes are blocked as the involved vehicles moved to the shoulder of the road after the accident. As shown in Figure 12(a), the first two events (i.e., Event 350 and Event 2116) happened at the beginning of morning and afternoon rush hours, and the last event (i.e., Event 2621) happened near the end of the afternoon rush hour. As illustrated in Figure 12(b), the prediction accuracy of H-ARIMA+ improves the prediction accuracy of H-ARIMA, NNet by up to 45% and 67%, respectively. We observe that though H-ARIMA can capture the sudden speed changes at rush hours, it cannot predict traffic in case of events. This is because the effect of traffic events are smoothed in historical averages.

We also study the effect of road construction events on our prediction model. Figure 13 shows the effect of a 6-hour long road construction event which happened in I-405 on a specific sensor. There is one lane affected by this event and the distance between this event and the selected sensor is 0.23 mile. As shown in Figure 13(a), the traffic speed deviates sharply especially in the first hour of the event.

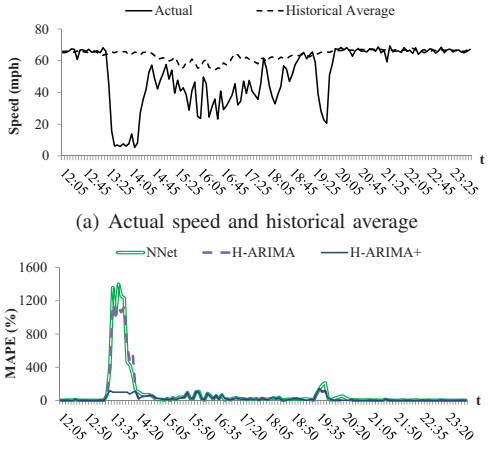


(a) Actual speed and historical average

(b) MAPE of the sensor

Figure 12. Case study on traffic collision events

Similar to traffic collision events, since ARIMA cannot handle sudden speed changes, and HAM cannot react to traffic dynamics such as events, the prediction accuracy of H-ARIMA (which selects between ARIMA and HAM) is very low at the beginning half an hour. However, H-ARIMA+ utilizes the event information, and yields significantly better prediction at the beginning of this event by improving H-ARIMA and NNet by up to 91% (see Figure 13(b)).



(a) Actual speed and historical average

(b) MAPE of the sensor

Figure 13. Case study on a road construction event

We summarize our findings in Table IV showing the overall precision of predictions on all sensors aggregated through all time stamps in terms of RMSE. As shown, H-ARIMA outperforms the baseline approaches in both prediction horizons. Moreover, when h=6 (i.e. 30 mins in advance prediction), H-ARIMA+ improves the accuracy of H-ARIMA by incorporating event information.

VII. CONCLUSION

In this paper we study a traffic prediction technique that uses real-world spatiotemporal traffic sensor data on road networks. We show that the traditional prediction approaches

Table IV
RMSE OF ALL SENSOR PREDICTION ON WEEKDAYS

	ES	ARIMA	NNet	H-ARIMA	H-ARIMA+
h=1	3.389	3.235	3.315	3.208	N/A
h=6	5.518	4.545	4.154	4.079	3.937

that treat traffic data streams as generic time series fail to forecast traffic during traffic peak hours and in the case of events such as accidents and road constructions. Our proposed algorithm significantly improves the prediction accuracy of existing approaches by incorporating the historical traffic data into the prediction model as well as correlating the event attributes with traffic congestion. In this paper, we studied the prediction problem for each sensor individually. In future, we plan to consider the spatial correlations between sensors to improve the prediction accuracy.

ACKNOWLEDGMENT

This research has been funded in part by NSF grant IIS-1115153, a contract with Los Angeles Metropolitan Transportation Authority (LA Metro), the USC Integrated Media Systems Center (IMSC), HP Labs and unrestricted cash gifts from Northrop Grumman and Microsoft. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of the sponsors such as the National Science Foundation or LA Metro.

REFERENCES

[1] http://www.nytimes.com/2011/05/13/technology/13data.html. last visited june 14, 2012.

[2] M. Ben-akiva, M. Bierlaire, H. Koutsopoulos, and R. Mishalani. DynaMIT: a simulation-based system for traffic prediction. In *DACCORD'98, Delft, The Netherlands*.

[3] G. Box and G. Jenkins. *Time series analysis: Forecasting and control.* San Francisco: Holden-Day, 1970.

[4] M. M. Chong, A. Abraham, and M. Paprzycki. Traffic accident analysis using decision trees and neural networks. In *IADIS'04, Portugal*.

[5] S. Clark. Traffic prediction using multivariate nonparametric regression. In *JTE'03*, volume 129.

[6] U. Demiryurek, F. Banaei-Kashani, C. Shahabi, , and A. Ranganathan. Online computation of fastest path in time-dependent spatial networks. In *SSTD'11*.

[7] J. D. Gehrke and J. Wojtusiak. A natural induction approach to traffic prediction for autonomous agent-based vehicle route planning. MLI 08-1, George Mason University.

[8] M. A. Hall and L. A. Smith. Practical feature subset selection for machine learning. In *ACSC98, Perth, Berlin: Springer.*, pages 181–191, 1998.

[9] A. Ihler, J. Hutchins, and P. Smyth. Adaptive event detection with time-varying poisson processes. In *ACM SIGKDD'06*, NY, USA.

[10] S. Ishak and C. Alecsandru. Optimizing traffic prediction performance of neural networks under various topological, input, and traffic condition settings. In *JTE'04*, volume 130.

[11] J. Kwon, M. Mauch, and P. P. Varaiya. Components of congestion : delay from incidents, special events, lane closures, weather, potential ramp metering gain, and excess demand. In *TRR'06*, pages 84–91.

[12] S. Lee and D. B. Fambro. Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting. In *TRR'98*.

[13] X. Li, Z. Li, J. Han, and J. Lee. Temporal outlier detection in vehicle traffic data. In *ICDE '09.*

[14] X. Li and F. Porikli. A hidden markov model framework for traffic event detection using video features. In *ICIP '04*, 2004.

[15] R. S. Marshment, R. C. Dauffenbach, and D. A. Penn. Short-range intercity traffic forecasting using econometric techniques. In *ITE Journal*, volume 66, 1996.

[16] N. L. Nihan and J. N. Zhu. Short-term forecasts of freeway traffic volumes and lane occupancies, phase 1. In *TNW'93*, volume 5.

[17] B. Park, C. J. Messer, and T. I. Urbanik. Short-term freeway traffic volume forecasting using radial basis function neural network. Number 1651.

[18] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[19] V. J. Stephanedes, P. G. Michalopoulos, and R. A. Plum. Improved estimation of traffic flow for real-time control discussion and closure. In *TRR'81*, number 795.

[20] C. Tsai and S. Tai. Traffic monitoring and event analysis at intersection based on integrated multi-video and petri net process. In *MMM'11*, Berlin, Heidelberg.

[21] J. van Lint, S. Hoogendoorn, and H. van Zuylen. Freeway travel time prediction with State-Space neural networks. In *TRR'02*, volume 1811.

[22] B. Williams, P. Durvasula, and D. Brown. Urban freeway traffic flow prediction: Application of seasonal autoregressive integrated moving average and exponential smoothing models. In *TRR'98*, volume 1644.

[23] H. Xiao, H. Sun, B. Ran, and Y. Oh. Fuzzy-Neural network traffic prediction framework with wavelet decomposition. In *TRR'03*, volume 1836.

[24] J. Yuan, Y. Zheng, X. Xie, and G. Sun. Driving with knowledge from the physical world. In *SIGKDD'11*.

[25] H. M. Zhang. Recursive prediction of traffic conditions with neural network models. In *JTE'00*, volume 126.