Traffic Flow Forecasting for Urban Work Zones

Yi Hou, Praveen Edara, *Member, IEEE*, and Carlos Sun

***Abstract*—None of numerous existing traffic flow forecasting models focus on work zones. Work zone events create conditions that are different from both normal operating conditions and incident conditions. In this paper, four models were developed for forecasting traffic flow for planned work zone events. The four models are random forest, regression tree, multilayer feedforward neural network, and nonparametric regression. Both long-term and short-term traffic flow forecasting applications were investigated. Long-term forecast involves forecasting 24 h in advance using historical traffic data, and short-term forecasts involves forecasting 1 h and 45, 30, and 15 min in advance using real-time temporal and spatial traffic data. Models were evaluated using data from work zone events on two types of roadways, a freeway, i.e., I-270, and a signalized arterial, i.e., MO-141, in St. Louis, MO, USA. The results showed that the random forest model yielded the most accurate long-term and short-term work zone traffic flow forecasts. For freeway data, the most influential variables were the latest interval's look-back traffic flows at the upstream, downstream, and current locations. For arterial data, the most influential variables were the traffic flows from the three look-back intervals at the current location only.**

***Index Terms*—Intelligent transportation system (ITS), neural network, nonparametric regression, random forest, regression tree, traffic flow forecasting, work zones.**

I. Background

A SIGNIFICANT type of congestion in urban areas is recurrent, occurring at bottlenecks where demand significantly exceeds capacity during peak hours of travel. Given the high frequency and severity of bottlenecks in major urban areas, it is easy to understand why most of the traffic flow forecasting research has focused on recurrent congestion. Nonrecurring causes of congestion such as those resulting from work zones, special events, or incidents are also significant contributors of congestion. According to the U.S. Department of Energy [1], work zones are accountable for approximately 24% of the nonrecurring delay experienced by motorists. In recent years, intelligent transportation systems (ITS) have been deployed in work zones to improve traffic flow and safety. Traffic Management Centers (TMCs) have deployed ITS technologies such as dynamic message signs, variable speed limits (VSLs), and queue warning systems. Accurate traffic flow forecasts are necessary for the scheduling and operation of work zones. Scheduling tools use traffic flow during different times of day as input in determining the least impactful times for closing lanes. Long-term forecasts are usually sufficient for work zone scheduling applications. The use of ITS in work zones for active traffic control, not just traveler information, has increased in recent years. An example is the use of VSL systems that change speed limits based on traffic flow inside and upstream from a work zone. Such a system operates in real time and relies on traffic flow data measured from traffic sensors. A proactive VSL system that can anticipate traffic flow conditions in the near future and adjust speed limits before the flow deteriorates is more valuable than a reactive system that relies on past flow values alone. Ramp metering, hard shoulder running, and other applications are also being deployed at work zones and can all benefit from accurate short-term traffic flow forecasts. There is an abundance of traffic flow forecasting models from previous research. However, few studies have focused on the impact of dynamic variation of demand and capacity resulting from work zones. The presence of a work zone and its characteristics such as the type of work, number of closed lanes, reduced lane width, and other variables not only affect roadway capacity but also travel demand. This underresearched topic of traffic flow forecasting for work zones is discussed in this paper. Four models were developed for forecasting traffic flow for planned work zone events. The four models were the following: multilayer feedforward neural network, nonparametric regression, regression tree, and random forest. Both long-term and short-term traffic flow forecasting were investigated. Long-term prediction allows TMC to provide traffic operation plan before work zone is deployed. The short-term prediction allows real-time traffic control in work zone. While long-term forecast was made 24 h in advance using historical traffic data, short-term forecasts were made 1 h and 45, 30, and 15 min in advance using real-time temporal and spatial traffic data. The proposed models were evaluated using urban work zone data on two types of roadways, a freeway, i.e., I-270, and a signalized arterial, i.e., MO-141, in St. Louis, MO, USA. These data were collected from June 2012 to September 2013.

Manuscript received April 23, 2014; revised September 1, 2014 and October 29, 2014; accepted November 12, 2014. Date of publication December 11, 2014; date of current version July 31, 2015. The Associate Editor for this paper was S. Sun.

The authors are with the Department of Civil Engineering, University of Missouri, Columbia, MO 65211 USA (e-mail: yhcz4@mail.missouri.edu; edarap@missouri.edu; csun@missouri.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2014.2371993

II. Literature

As already mentioned, existing research on traffic flow forecasting primarily focused on non-work-zone conditions. A review of this literature and the rationale for the four models chosen for work zone flow forecasting are presented here. Time series analyses have been traditionally used for forecasting traffic flows. A great majority of time series approaches are univariate in nature. Univariate time series models use only historical traffic flow data from the location of interest to predict future traffic flow at the same location. The family of autoregressive integrated moving average (ARIMA) models was the

most extensively applied time series model form [2], [3], [4]. Other time series models included nonparametric regression [5], [6], [7], local linear regression [8], and Kalman filtering [9], [10].

Multivariate time series models were later developed to take into account both temporal and spatial traffic flow information into forecasting. Williams [11] developed a multivariate ARIMA model that included upstream traffic flow data. Stathopoulos and Karlaftis [12] proposed a multivariate time series state-space model that used traffic flow data from an upstream detector and found that the model resulted in higher prediction accuracy than univariate time series models. Kamarianakis and Prastacos [13] proposed the space-time ARIMA models to predict traffic flow in an urban area.

Another major class of traffic flow forecasting models was the artificial neural network. A large number of neural networks ranging from static to dynamic structures were deployed, including multilayer feedforward [14]–[19], radial basis function [20], [21], time delayed [22]–[26], and recurrent [27], [28].

Recently, Sun *et al.* [29] and Sun and Xu [30] have proposed Bayesian network approaches to forecasting traffic flow on a link using spatial traffic flow data from adjacent road links. Zhang and Ye [31] proposed a fuzzy logic system to improve traffic flow prediction accuracy. Min and Wynter [32] developed an extended time-series-based method that incorporated temporal and spatial interactions. Pan *et al.* [33] used a stochastic cell transmission framework to predict short-term traffic flow by considering the spatial–temporal correlation in the network. Sun *et al.* [34] conducted research on network-scale traffic modeling and forecasting with graphic lasso and neural networks. Huang and Sun [35] applied kernel regression with sparse metric learning to forecast short-term traffic flow.

From the reviewed literature, nearest neighbor nonparametric regression [5], [6], [7] and multilayer feedforward neural network [14]–[19] models were selected for application to work zone flow forecasting. Two other models that have not been used in previous flow forecasting research, i.e., regression tree and random forest, were proposed for the first time here. Regression tree [36] has emerged as one of the most popular methods for both classification and regression. By performing a binary split on input variables, it can classify data into patterns. It can also simultaneously treat a combination of numeric and categorical variables. A regression tree's intuitive model structure allows for robust interpretation of results. Random forest [37] is a powerful tool for data mining and machine learning and has been implemented in a variety of disciplines. It is an ensemble learning method that combines a large group of regression trees. The main idea behind ensemble methods is to build a strong prediction model by combining a large group of weak models. Unlike some data mining methods that lack good model interpretation capabilities because they are seen as "black boxes," the prediction process of random forest can be intuitively interpreted by estimating predictor importance. Random forest is robust to noise and outliers, a common characteristic in traffic flow data obtained from traffic sensors.

The remainder of this paper is organized as follows. In Section III, the methodology is introduced. The fundamental theory behind the four methods is offered. The measures of effectiveness to compare the performance of the methods are also presented. In Section IV, data collection, experimental design, and the study findings are described. Both long-term and short-term forecasting applications are discussed. The conclusion is drawn in the final section.

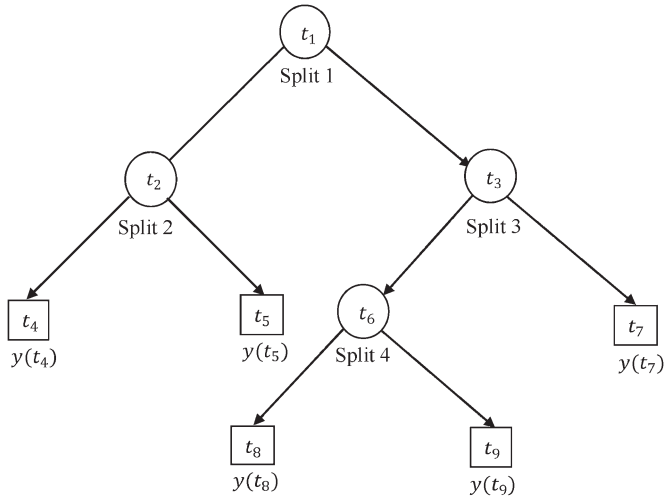


Fig. 1. Regression tree structure.

III. Methodology

Here, the following notations are used to describe the four methods used in this paper: h: hour of day; d: day of the week; m: month of the year; wt: type of work zone; l_{total}: total number of lanes; l_{closed}: number of lanes closed; sl: work zone speed limit; bl: work zone begin log on roadways; dur: work zone duration; and len: work zone length.

A. Regression Tree and Random Forest

A regression tree is similar to a classification tree or a decision tree [36]. It is constructed by a sequence of binary splits of training set X into terminal nodes, as shown in Fig. 1. Each terminal node is assigned an output $y(t)$, the predicted traffic flow.

The tree construction process revolves around three steps: 1) the selection of the splitting rules; 2) the criterion to stop splitting and declare a terminal node; 3) the assignment of a value $y(t)$ to each terminal node. At each node split, a question, "Is $x_i \geq a$?," is asked, where x_i is one of the predictor variables in $\{h, d, l_{\text{total}}, m, \text{wt}, l_{\text{closed}}s, l, \text{bl}, \text{dur}, \text{len}\}$, and a is a threshold value. The binary answer to the question is used to split the node. The purpose of the binary split at each node in a regression tree is to reduce the overall resubstitution estimate of prediction error. Resubstitution is synonymous with training. The resubstitution estimate, i.e., $R(t)$, of prediction error of node t is often measured by the mean squared error of the node subset X_t. It is formulated as [36]

$$R(t) = \frac{1}{N_t} \sum_{i=1}^{N_t} (y_{ti} - \bar{y}_t)^2 \tag{1}$$

where N_t is the size of node t, y_{ti} is the traffic flow of the ith data point falling into node t, and $\bar{y}_t$ is the average of traffic flows of subset falling into node t. The best predictor variables x^* and threshold value a^* that split node t are selected by decreasing $R(t)$ the most. For any split of node t into two descendant nodes t_Y and t_N, the decrease of $R(t)$ is

$$\Delta R(t) = R(t) - R(t_L) - R(t_R). \tag{2}$$

By adopting a stop-splitting rule, a node is declared as the terminal node if node size $N_t \leq N_{\min}$. Breiman *et al.* [36] suggested $N_{\min}$ for regression tree to be 5. The node assignment value, i.e., $y(t)$, can be easily obtained by averaging the response values of the terminal subset X_t.

A drawback associated with regression tree is high variance. Bagging or bootstrap aggregating is an approach to reduce the variance of an estimated prediction function. The essential idea of bagging is to build a large group of regression trees and average the results produced by each tree. Bagging performs particularly well for high-variance and low-bias procedures. Regression trees generated in bagging are identically distributed. Suppose K trees are generated in bagging with positive pairwise correlation ρ, and each with variance σ^2. The variance of the average is

$$\text{Var}(\mu) = \rho\sigma^2 + \frac{1-\rho}{K}\sigma^2. \tag{3}$$

With the increase in$K$, the second term approaches zero, but the first term remains, and the pairwise correlation of bagged trees limits the benefit of averaging. In order to reduce the correlation between trees, random selection of the predictor variables is performed at each node split during the tree growing process. A combination of regression trees built through this process is called a random forest [37].

The random forest algorithm for regression applications is described as [38] follows.

1. For $k = 1$ to K:
   a. Draw a bootstrap sample from the original training data set.
   b. Grow a regression tree T_k with the bootstrapped data set by recursively repeating the following steps for each node until the minimum node size that is set by user is reached.
      i. Randomly select m predictors from $\{h, d, l_{\text{total}}, m, \text{wt}, l_{\text{closed}}, \text{sl}, \text{bl}, \text{dur}, \text{len}\}$.
      ii. By using the same node splitting rule that is introduced in regression tree, select the predictor variable among m predictor variables that makes the best split to split the node into two descendant nodes.
2. Output a collection of trees $\{T_k\}_K^1$.
3. To make a prediction for a new data point x

$$\hat{f}_{rf}^K(x) = \frac{1}{K}\sum_{k=1}^{K} T_k(x). \tag{4}$$

To predict a test data case, data are pushed down all the regression trees. Each tree will produce a predicted traffic flow. The end result is the average of the predicted traffic flows of all trees. Breiman [37] recommended using $m = p/3$ and minimum node size of 5 for regression applications. He presented two reasons for using bagging. First, the use of random features improves the accuracy. Second, bagging allows for the continuous estimation of error of the "combined ensemble of trees" [37]. The error estimate is defined as the "error rate of the out-of-bag (OOB) classifier on the training set" [37]. The OOB estimate is similar to the N-fold cross-validation [36]. Thus, random forest can be fit in one sequence, with cross-validation being performed along the way. The training can be terminated after the OOB error stabilizes.

Random forest constructs a measure of variable importance to help the user understand the mechanism of the prediction process and to eliminate less important variables. The measure is obtained by accumulating improvement in split criterion at each split in each tree separately for each variable over all trees in the forest [38].

The computational complexity of random forest was examined by the standard order notation. Suppose the training data contain n data points and m attributes, according to [39], the computational cost of building one regression tree is $O(mn \log n)$. If K regression trees are built in a random forest, then the computational complexity of the random forest is $O(K(mn \log n))$.

B. Multilayer Feedforward Neural Network

A neural network uses linear combinations of input variables, $\{h, d, l_{\text{total}}, m, \text{wt}, l_{\text{closed}}, \text{sl}, \text{bl}, \text{dur}, \text{len}\}$, to derive features that are then combined using a nonlinear function to predict the traffic flow [38]. A multilayer feedforward neural network consists of an input layer, i.e., X_k, an output layer, i.e., Y, and one or more hidden layers, i.e., Z_m. The linear combinations of input variables create new features Z_m that form a hidden layer, since the features are not actually observed. The traffic flow is modeled as a function of linear combinations of Z_m. Equations (5) and (6) show how the features and the predicted flow are calculated [38], i.e.,

$$Z_m = \sigma\left(\alpha_{0m} + \alpha_m^T X\right), \qquad m = 1, \ldots, M \tag{5}$$

$$Y = g(\beta_0 + \beta^T Z) \tag{6}$$

where $X = \{h, d, l_{\text{total}}, m, \text{wt}, l_{\text{closed}}, \text{sl}, \text{bl}, \text{dur}, \text{len}\}$, and $Z = \{Z_1, Z_2, \ldots, Z_M\}$. The activation function $\sigma(x)$ is chosen as sigmoid $\sigma(x) = 1/(1 + \exp(-x))$ for this paper, since it is a commonly used activation function [17], [18]. For regression, the output function $g(x)$ is usually chosen as the identity function $g(x) = x$.

C. Nearest Neighbor Nonparametric Regression

Nonparametric regression describes the relationship between the independent variables and the traffic flow to be predicted, based on observed data. The nearest neighbor nonparametric regression is a pattern matching method that matches the current observations with those in a database of historical observations.

Given a distance measure and a forecasting procedure, the model is presented as follows.

- Identify the k nearest neighbors from the training data set for an unknown data point.
- Using the traffic flows of the selected neighbors, forecast the traffic flow of the unknown data point.

For this paper, a common distance measure—Euclidean distance—is used to identify k nearest neighbors. It is formulated as

$$D(\boldsymbol{x_j}, \boldsymbol{x_k}) = \sqrt{\sum_{i=1}^{l}(x_{ji} - x_{ki})^2} \tag{7}$$

where x_j and x_k are independent variables of two data points. The forecast of traffic flow for the unknown data point can be calculated by the weighted inverse of distance as follows [6]:

$$\hat{y} = \left(\sum_{i=1}^{k} \frac{y_i}{D_i} \Big/ \sum_{i=1}^{k} \frac{1}{D_i}\right) \tag{8}$$

where $\hat{y}$ is the predicted traffic flow, y_i is the traffic flow of the ith nearest neighbor, and D_i is the distance between the ith nearest neighbor and the unknown data point.

D. Measures of Effectiveness

Performance of machine learning methods is typically evaluated using error measures. The four models were evaluated using three common measures of effectiveness: root-mean-square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). They are formulated as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(v_i - \widehat{v_l})^2} \tag{9}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|v_i - \widehat{v_l}| \tag{10}$$

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|v_i - \widehat{v_l}|}{v_i} \tag{11}$$

where v_i is the observed traffic flow, $\widehat{v_l}$ is the predicted or forecasted traffic flow, and N is the total number of observations.

IV. Experimental Design and Results

A. Data Collection

In this paper, detailed urban work zone and traffic flow data were collected on two different types of roadways in St. Louis. One is a segment of I-270 freeway between MO-370 and MO-367, and the other is a segment of MO-141 between Clayton Rd. and I-55, an urban signalized arterial. I-270 is the busiest freeway in the metropolitan St. Louis area with up to 180 000 vehicles per day. The length of the I-270 segment

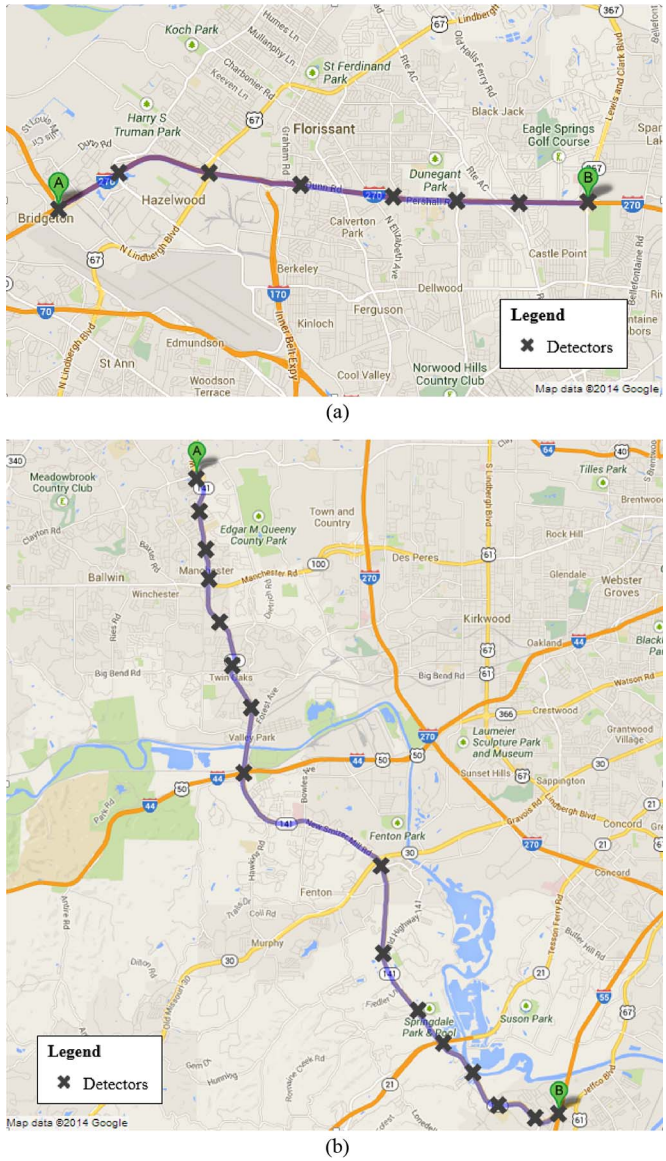


Fig. 2. I-270 and MO-141 study segments. (a) Study segment of I-270. (b) Study segment of MO-141.

is approximately 12 mi, and the free-flow travel time along this segment is approximately 11 min. Eight detectors are located along this stretch of I-270. The segment of MO-141 is approximately 17 mi long. Free-flow travel time for this stretch is approximately 23 min. There are 16 detectors along the segment. Fig. 2 depicts the geometric characteristics and locations of detectors for both roadway segments. Traffic flow data were collected in half-minute resolutions by detectors 24 h a day for 15 months from June 2012 to September 2013. A total of 69 work zones on I-270 and 92 work zones on MO-141 were deployed during the data collection period. Detailed work zone data such as work zone type, number of lanes closed, speed limit, starting location of work zone, duration, and work zone length were retrieved from Missouri Department of Transportation database. Since the goal of this paper was to predict work zone traffic demand, the traffic flows located between 1 and 3 mi upstream of work zone taper were used as dependent variables for model development.

TABLE I
SUMMARY OF LONG-TERM PREDICTION RESULTS FOR THE (a) I-270 AND (b) MO-141 DATA SETS

(a)

	RMSE	MAE	MAPE
Baseline Predictor	312.2	149.6	9.37%
Regression Tree	279.7	143.1	8.98%
Random Forest	**231.5**	**119.8**	**7.70%**
Neural Network	235.9	141.0	8.95%

Note: The prediction accuracy for baseline predictor only shows those that can be predicted. 1042 out of 3957 observations could not be predicted.

(b)

	RMSE	MAE	MAPE
Baseline Predictor	140.6	73.6	17.95%
Regression Tree	152.1	77.2	17.57%
Random Forest	**102.4**	**57.0**	**14.06%**
Neural Network	111.0	68.6	21.07%

Note: The prediction accuracy for baseline predictor only shows those that can be predicted. 1146 out of 2341 observations could not be predicted.

B. Long-Term Traffic Demand Prediction

Multilayer feedforward neural network, regression tree, and random forest were implemented for long-term traffic prediction. The nearest neighbor nonparametric regression was not used for long-term prediction since the data consisted of several categorical variables. Long-term prediction provides prediction for a time horizon of 24 h based on historical traffic data. Eleven variables that are relevant to work zone traffic demand were chosen as predictors for the long-term traffic demand models. They are hour of the day, day of the week, month of the year, total number of lanes, number of closed lanes, work zone type, speed limit, direction, work zone begin log, work zone duration, and work zone length. Some of these variables, such as the work zone length and number of closed lanes, are commonly used in work zone capacity estimation models [40]. Traffic flows were aggregated into 1-h intervals, since hourly flow is the most commonly used traffic flow parameter. For both the I-270 and MO-141 data sets, 60% of the data were randomly selected as training data and the rest as testing data. Separate models were built for freeway (I-270) and signalized urban arterial (MO-141).

Using the experience of other researchers along with trial-and-error processes, the best values for model parameters were determined for all proposed models. For the multilayer feedforward neural network, network structures of single and multiple hidden layers were tried. Backpropagation algorithm with Levenberg–Marquardt optimization [41] was used to train the model. The initial scalar μ is 0.001, and the decrease and increase factors of μ are 0.1 and 10. After varying the number of hidden layers and the number of nodes, a network structure of two hidden layers with 40 nodes in each produced the best performance for the I-270 data set. A network structure of one hidden layer with 60 nodes produced the best results for the MO-141 data set. The regression tree yielded the best model

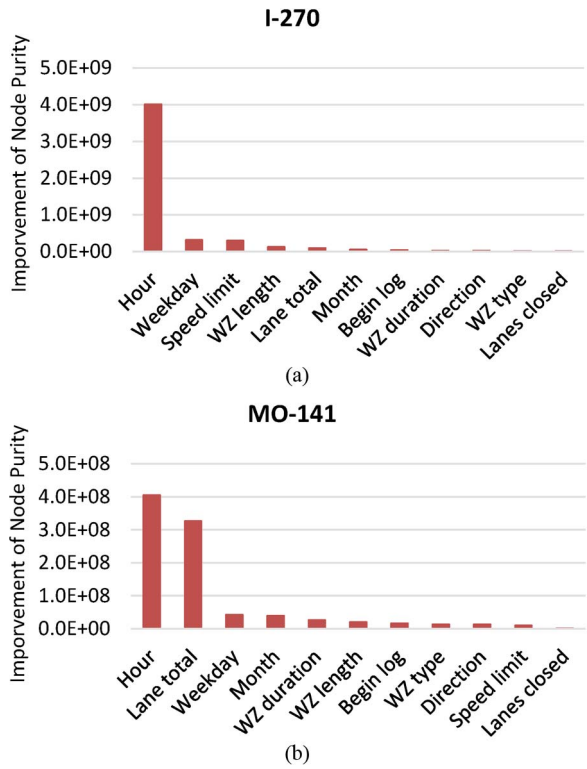


Fig. 3. Variable importance in long-term traffic demand prediction.

performance for both data sets when the minimum terminal node size of 5 was used. In the process of developing random forest, the OOB error estimate stabilized after approximately 100 trees. The minimum terminal node size of 5 and the number of randomly selected input variables of 8 resulted in the best model performance for both data sets. Historical average traffic demand was used for comparison as a baseline. The historical averages were conditioned on hour of the day, day of the week, month of the year, total number of lanes, number of closed lanes, work zone type, and speed limit. There were instances when the work zone characteristics in the testing sample were not observed in historical training data. Thus, baseline predictions for such unobserved work zone characteristics were not made.

Table I summarizes the results from the test data for both I-270 and MO-141. Values in bold typeface indicate the smallest values for RMSE, MAE, and MAPE. As shown in Table I, random forest outperformed regression tree, multilayer feedforward neural network, and baseline predictor for all three measures. As expected, the ensemble method of bagging of random forest improved model prediction accuracy over the regression tree. The importance of random forest variables is presented in Fig. 3. The Y-axis in the figure is the accumulated improvement of node purity at each split in each tree of a variable. A high value of improvement indicates high variable importance. Fig. 3 reveals that hour of day dominates traffic prediction for freeway work zone. This is presumably due to the significant variation of traffic flow during the day. Among the work zone characteristics, work zone speed limit and length of work zone exhibited the highest influence. The time of day variable was once again the most influential for the MO-141 arterial corridor. Total number of lanes also had a very high influence

on the long-term forecasts. Work zone duration and length were the most influential among all work zone characteristics.

C. Short-Term Traffic Demand Prediction

Multilayer feedforward neural network, nearest neighbor nonparametric regression, regression tree, and random forest models were developed for short-term traffic prediction of work zones on the two routes. Let $V(t+i)$ denote the predicted flow for the $(i+1)$th time interval in the future. For the first prediction interval $(i=0)$, the flow is denoted by $V(t)$. Let $V(t-j)$ denote the observed flow for the jth time interval in the past. For example, $V(t-1)$ denotes the flow for the most recent time interval in the past.

The future traffic flows at the location, i.e., $V(t)$, $V(t+1)$, …, $V(t+b-1)$, are likely correlated to the traffic flows in the recent past at the subject location, i.e., $V(t-1)$, $V(t-2)$, …, $V(t-d)$, the adjacent upstream location, i.e., $V_{\mathrm{us}}(t-1)$, $V_{\mathrm{us}}(t-2), \ldots, V_{\mathrm{us}}(t-d)$, and the adjacent downstream location, i.e., $V_{\mathrm{ds}}(t-1)$, $V_{\mathrm{ds}}(t-2), \ldots, V_{\mathrm{ds}}(t-d)$, where b and d parameters denote the extent of prediction and "look-back" intervals. The short-term traffic demand prediction problem essentially predicts the values of $V(t)$, $V(t+1)$, …, $V(t+b-1)$, using the time series flows from the detector of interest and adjacent upstream and downstream detectors. In addition to temporal and spatial traffic flow information, other relevant factors such as hour of day, day of week, number of lanes, speed limit, direction, and work zone begin log were also considered in the models. Because traffic flows at more distant locations have less influence, only observations taken from within 2 mi of the detector of interest were used in the models. Traffic flows were aggregated into 15-min intervals, considering that 15 min is a reasonable amount of time for a TMC to implement a traffic control strategy and for drivers to react to the new strategy. The extent of prediction interval was set to 4 ($b=4$). In other words, the proposed models provide prediction of traffic demand over 15-min intervals for up to 1 h in advance.

Prediction on multiple time periods into the future enables a wider range of applications. For signalized arterials, forecasting 1-min traffic flow may be more useful for signal timing. Thus, for the MO-141 arterial segment, forecasts were also made using 1-min intervals (in addition to the 15-min forecasts). The extent of look-back was varied from one to six intervals; adding more look-back intervals after 3 did not improve model performance. Thus, the extent of look-back interval was set to 3. For both the I-270 and MO-141 data sets, 60% of data were randomly selected for training and the rest for testing.

Multilayer feedforward neural networks with single and multiple hidden layers were tried. Backpropagation algorithm with Levenberg–Marquardt optimization [41] with the same parameters used for long-term forecasting was used to train the model. A network of two hidden layers with 40 nodes in the first hidden layer and 20 nodes in the second produced the best results for I-270, whereas two hidden layers with 40 nodes in the first hidden layer and ten nodes in the second produced the best results for MO-141. The best results for the nonparametric model for I-270 and MO-141 were obtained when the numbers of nearest neighbors were 9 and 7, respectively. A minimum terminal node size of 5 yielded the best regression tree results. For random forest, after 100 trees were built, further increases in the number of trees did not improve model performance. The minimum terminal node size of 5 and the number of randomly selected input variables being 6 resulted in the best model performance for both data sets. Instantaneous traffic demand was used as a baseline predictor for comparison. The instantaneous traffic demand for short-term prediction is to use the last measured traffic demand as a proxy for the traffic demand of future intervals. Instantaneous traffic demand provides accurate predictions in cases where traffic conditions change slowly over long time periods.

TABLE II
SUMMARY OF 15-MIN TRAFFIC FLOW PREDICTION RESULTS FOR THE (a) I-270 AND (b) MO-141 DATA SETS

(a)

		t	$t+1$	$t+2$	$t+3$
Baseline Predictor	RMSE	54.7	79.9	102.9	126.0
	MAE	38.2	53.5	69.5	85.9
	MAPE	10.62%	14.69%	19.26%	24.07%
Regression Tree	RMSE	50.2	57.4	59.8	63.2
	MAE	34.1	36.3	37.0	38.9
	MAPE	9.62%	10.42%	10.88%	11.25%
Random Forest	RMSE	**35.5**	**39.8**	**40.1**	**43.5**
	MAE	**23.9**	**25**	**26.1**	**27.3**
	MAPE	**6.85%**	**7.25%**	**7.71%**	**7.96%**
Neural Network	RMSE	39.7	44.4	50.0	53.2
	MAE	26.8	29.2	31.4	32.8
	MAPE	7.79%	8.68%	9.35%	9.68%
Nearest Neighbor	RMSE	39.8	49.8	57.3	64.4
	MAE	25.7	30.8	35.3	40
	MAPE	7.47%	9.13%	10.89%	12.74%

(b)

		t	$t+1$	$t+2$	$t+3$
Baseline Predictor	RMSE	26.9	40.9	54.3	67.5
	MAE	14.5	23.1	31.0	39.1
	MAPE	9.17%	16.24%	22.24%	29.21%
Regression Tree	RMSE	27.5	32.4	37.5	38.8
	MAE	14.2	17.9	19.7	20.1
	MAPE	8.33%	11.98%	12.92%	13.48%
Random Forest	RMSE	**20.5**	**23.1**	**25.5**	**28.3**
	MAE	**10.6**	**13**	**14.1**	**15**
	MAPE	**6.64%**	**9.36%**	**10.18%**	**10.88%**
Neural Network	RMSE	21.0	26.2	29.2	34.5
	MAE	11.8	15.3	17.2	19.8
	MAPE	8.32%	11.72%	13.15%	15.11%
Nearest Neighbor	RMSE	23.5	30.5	38.5	45.8
	MAE	12.7	16.8	20.9	24.9
	MAPE	8.24%	11.96%	14.87%	17.70%

Table II presents the prediction results from the test data for both I-270 and MO-141. In the table, each "t," "$t+1$," "$t+2$," and "$t+3$" is a 15-min interval into the future, i.e., 15-, 30-, 45-, and 60-min prediction intervals. The performance

TABLE III
SUMMARY OF 1-MIN TRAFFIC FLOW PREDICTION RESULTS FOR THE MO-141 DATA SET

		t	$t+1$	$t+2$	$t+3$
Baseline Predictor	RMSE	8.9	9.6	14.8	15.5
	MAE	4.2	4.5	6.6	7.5
	MAPE	13.47%	15.20%	18.70%	22.22%
Regression Tree	RMSE	8.6	9.1	9.6	10.1
	MAE	3.9	4.2	4.4	4.7
	MAPE	12.49%	13.62%	14.36%	15.47%
Random Forest	RMSE	**6.4**	**6.7**	**6.9**	**7.4**
	MAE	**3.0**	**3.1**	**3.3**	**3.5**
	MAPE	**9.79%**	**10.51%**	**11.28%**	**11.94%**
Neural Network	RMSE	6.5	6.8	7.1	7.6
	MAE	3.2	3.4	3.6	3.9
	MAPE	12.28	12.87	14.90	15.51
Nearest Neighbor	RMSE	6.8	7.2	7.7	7.9
	MAE	3.4	3.5	3.7	3.9
	MAPE	13.05%	13.25%	14.11%	14.58%

TABLE IV
COMPUTATION TIMES

	Model Construction	Prediction for Test Data	Total
Regression Tree	0.62s	0.03s	0.65s
Random Forest	41.48s	0.29s	42.17s
Neural Network	111.96s	0.70s	112.66s
Nearest Neighbor	NA	158.08s	158.08s

of all five predictors worsens when predicting further into the future. Values in bold typeface indicate the smallest values for RMSE, MAE, and MAPE. Table II shows that all the error measures for random forest are smaller than those for the other models for all four 15-min prediction intervals. The performance of the baseline predictor worsens by a factor of greater than 2 between t and $t+3$, whereas the random forest predictor MAPE increased slightly over 1% (I-270) and 4% (MO-141) between t and $t+3$. All five predictors performed worse on MO-141. The comparison between random forest and regression tree error measures shows that the ensemble method of bagging can increase prediction accuracy for the weak model. Table III presents the 1-min traffic flow prediction results for MO-141. Table III shows that all the error measures for random forest are again smaller than those for the other models for all four 1-min prediction intervals.

Since the main application of short-term traffic flow prediction is for real-time traffic operation and control, the average computation time for model construction and prediction for test data are important.

Table IV presents the computation time for the four models on an Intel core i3 central processing unit with 4-GB random access memory. With the exception of the nearest neighbor model, all other models provide predictions within 1 s. The total computation time for random forest is lower than that for multilayer feedforward neural network and nearest neighbor nonparametric regression, but higher than that for regression tree.

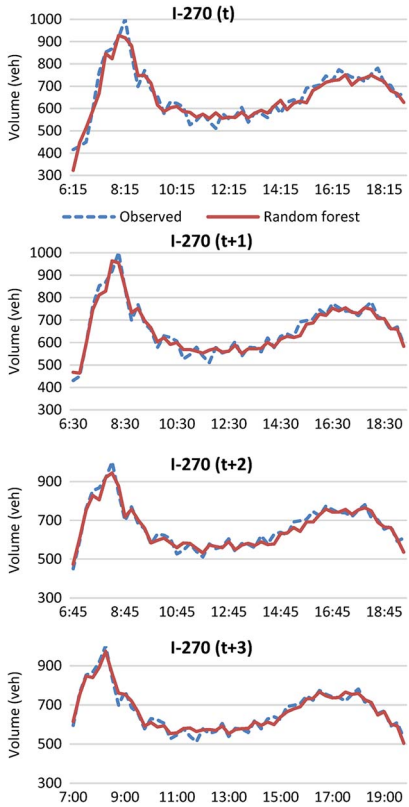


Fig. 4. Temporal variation of predicted and observed traffic demand on I-270 for short-term prediction.

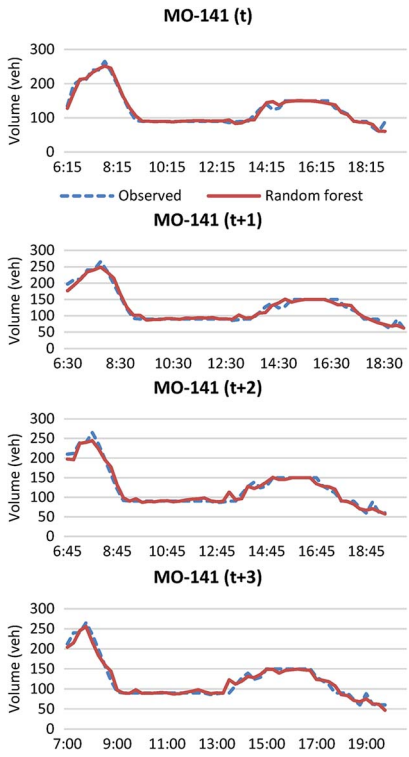


Fig. 5. Temporal variation of predicted and observed traffic demand on MO-141 for short-term prediction.

The work zone traffic demand of a randomly selected weekday on each study segment was forecasted by the best performed model. None of the observations on the randomly selected weekday was used in the model training process.

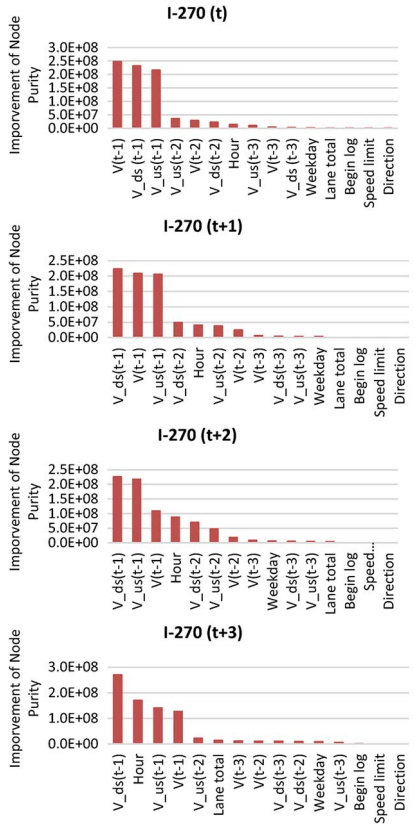


Fig. 6. Variable importance in short-term prediction for the I-270 data set.

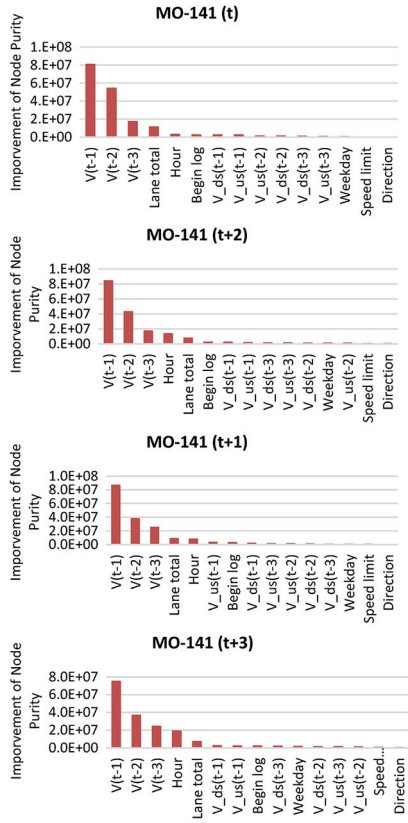


Fig. 7. Variable importance in short-term prediction for the MO-141 data set.

Figs. 4 and 5 present the temporal variation of predicted and observed traffic demand from 6:00 A.M. to 8:00 P.M. on a weekday on both I-270 and MO-141, respectively. Both figures show that the predicted values of random forest and the observed values are in close agreement for all four prediction time intervals. The random forest's MAPE values for this day for t, $t+1$, $t+2$, and $t+3$ prediction intervals were 4.34%, 2.63%, 2.92%, and 2.70% for I-270 and 4.20%, 5.39%, 4.89%, and 5.40% for MO-141.

The importance of variables was explored for short-term prediction. Figs. 6 and 7 display the importance of variables in predictions of all four time intervals for the I-270 and MO-141 data sets, respectively. Fig. 6 reveals that the most relevant variables in short-term traffic prediction for the I-270 data set are the flows at one look-back interval at the location of interest and the adjacent upstream and downstream locations, i.e., $V(t-1)$, $V_{\mathrm{us}}(t-1)$, and $V_{\mathrm{ds}}(t-1)$. As the forecast interval size increased from t to $t+3$, the importance of $V(t-1)$ decreased, whereas the importance of hour of day variable (*Hour*) increased. When forecasting flow for immediately next time interval (i.e., t), the previous interval's $(t-1)$ flow served as a good proxy for the hour of day variation in flow. As the forecast interval size increased (to $t+1$ and higher), this was not the case since the previous interval became more removed from the forecast interval, and thus, the hour of day variable started to exhibit a higher influence as the forecast interval size grew. The downstream flow, i.e., $V_{\mathrm{ds}}(t-1)$, became the primary variable onward, which might be indicative of congestion spillback occurring due to downstream bottlenecks that were felt at the current location in the future time intervals.

Fig. 7 shows that the most relevant variables in short-term traffic prediction for the MO-141 data set are the flows of all three look-back intervals at the location of interest, i.e., $V(t-1)$, $V(t-2)$, and $V(t-3)$. Unlike the I-270 short-term prediction, the variables exhibiting the highest influence did not vary with the size of prediction interval. It appears that signalization reduced the influence of upstream and downstream traffic flow values.

V. Conclusion

The lack of forecasting models for work zone traffic flow motivated the research presented in this paper. To that end, four models were developed for short-term and long-term traffic flow forecasting for work zones. Two of the models, i.e., regression tree and random forest, have not been investigated in previous traffic flow forecasting research. All three measures of effectiveness showed that random forest outperformed all the other models for both short-term and long-term forecasts.

Of all variables, the hour of day of the work zone was found to have the greatest impact in the long-term flow prediction for both freeway and signalized urban arterial. Short-term flow forecasts for freeway work zones depended the most on the flow values at the location of interest and at upstream and downstream locations during the latest look-back interval, i.e., $V(t-1)$, $V_{\mathrm{us}}(t-1)$, and $V_{\mathrm{ds}}(t-1)$. In contrast, the short-term flow forecast for the arterial work zones relied the most on the flow in the previous three look-back intervals only at the location of interest, i.e., $V(t-1)$, $V(t-2)$, and $V(t-3)$. The results demonstrated that the developed work zone traffic

flow forecast models did not find the work zone characteristics to play as significant a role as traffic flow variables. Recall that the training and testing samples included only work zone data. The results, however, do not mean that separate models are not needed for work zones. The relative importance of the traffic flow variables, time of day, day of the week, and other seasonality variables could be very different between normal operations and work zones. One possible reason for work zone characteristics not playing a significant role in the prediction is that the traffic flow variables serve as a surrogate for different work zone characteristics. For example, intensity of work activity and number of closed lanes affect traffic flow. Thus, by including traffic flow as a surrogate accounts for the combined influence of various work zone characteristics.

A few related research topics can be explored in the future. First, forecasting models by the type of work zone may be developed. For example, maintenance and construction work activities involve different intensities and durations. Thus, separate models for these two types may be warranted. Second, the proposed models may be applied to other urban areas to further investigate forecasting models. Third, other artificial intelligence and advanced time-series models may be explored in the future.

REFERENCES

[1] "Temporary losses of highway capacity and impacts on performance," Oak Ridge National Laboratory, Tech. Rep. ORNL/TM-2002/3, 2002.

[2] M. Levin and Y. D. Tsao, "On forecasting freeway occupancies and volumes," *Transp. Res. Record*, vol. 773, pp. 47–49, 1980.

[3] M. M. Hamed, H. R. Al-Masaeid, and Z. M. Bani Said, "Short-term prediction of traffic volume in urban arterials," *J. Transp. Eng.*, vol. 121, no. 3, pp. 249–254, May 1995.

[4] B. M. Williams, P. K. Durvasula, and D. E. Brown, "Urban traffic flow prediction: Application of seasonal autoregressive integrated moving average and exponential smoothing models," *Transp. Res. Rec.*, vol. 1644, pp. 132–144, 1998.

[5] B. L. Smith and M. J. Demetsky, "Traffic flow forecasting: Comparison of modeling approaches," *J. Transp. Eng.*, vol. 123, no. 4, pp. 261–266, Jul. 1997.

[6] B. L. Smith, B. M. Williams, and K. R. Oswald, "Comparison of parametric and non-parametric models for traffic flow forecasting," *Transp. Res. Part C, Emerging Technol.*, vol. 10, no. 4, pp. 303–321, Aug. 2002.

[7] S. Clark, "Traffic prediction using multivariate nonparametric regression," *J. Transp. Eng.*, vol. 129, no. 2, pp. 161–168, Mar. 2003.

[8] H. Sun, H. X. Liu, H. Xiao, R. R. He, and B. Ran, "Short-term traffic forecasting using the local linear regression model," presented at the 82nd Annual Meeting Transportation Research Board, Washington, DC, USA, 2003.

[9] I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through Kalman filtering theory," *Transp. Res. Part B, Methodological*, vol. 18, no. 1, pp. 1–11, Feb. 1984.

[10] J. Whittaker, S. Garside, and K. Lindeveld, "Tracking and predicting network traffic process," *Int. J. Forecast.*, vol. 13, no. 1, pp. 51–61, Mar. 1997.

[11] B. M. Williams, "Multivariate vehicular traffic flow prediction: An evaluation of ARIMAX modeling," *Transp. Res. Rec.*, vol. 1776, pp. 194–200, 2001.

[12] A. Stathopoulos and M. G. Karlaftis, "A multivariate state-space approach for urban traffic flow modeling and prediction," *Transp. Res. Part C, Emerging Technol.*, vol. 11, no. 2, pp. 121–135, Apr. 2003.

[13] Y. Kamarianakis and P. Prastacos, "Space-time modeling of traffic flow," *Comput. Geosci.*, vol. 31, no. 2, pp. 119–133, Mar. 2005.

[14] B. L. Smith and M. J. Demetsky, "Short-term traffic flow prediction: Neural network approach," *Transp. Res. Rec.*, vol. 1453, pp. 98–104, 1994.

[15] M. S. Dougherty and M. R. Cobbet, "Short-term inter-urban traffic forecasts using neural networks," *Int. J. Forecast.*, vol. 13, no. 1, pp. 21–31, Mar. 1997.

[16] C. Ledoux, "An urban traffic flow model integrating neural networks," *Transp. Res. Part C, Emerging Technol.*, vol. 5, no. 5, pp. 287–300, Oct. 1997.

[17] S. Innamaa, "Short-term prediction of traffic situation using MLP-Neural Networks," presented at the 7th World Congress Intelligent Transportation Systems, Turin, Italy, 2000.

[18] H. M. Zhang, "Recursive prediction of traffic conditions with neural networks," *J. Transp. Eng.*, vol. 126, no. 6, pp. 472–481, Nov./Dec. 2000.

[19] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Optimized and meta-optimized neural networks for short-term traffic flow prediction: A genetic approach," *Transp. Res. C, Emerging Technol.*, vol. 13, no. 3, pp. 211–234, Jun. 2005.

[20] H. Chen, S. Grant-Muller, L. Mussone, and F. Montgomery, "A study of hybrid neural network approaches and the effects of missing data on traffic forecasting," *Neural Comput. Appl.*, vol. 10, no. 3, pp. 277–286, Dec. 2001.

[21] L. Chen and C. L. P. Chen, "Ensemble learning approach for freeway short-term traffic flow prediction," in *Proc. IEEE Int. Conf. Syst. Syst. Eng.*, San Antonio, TX, USA, 2007, pp. 1–6.

[22] P. Lingras and P. Mountford, "Time delay neural networks designed Using genetic algorithms for short-term inter-city traffic forecasting," in *Proc. IEA/AIE LNAI*, 2001, vol. 2070, pp. 290–299.

[23] S. Y. Yun, S. Namkoong, J. H. Rho, S. W. Shin, and J. U. Choi, "A performance evaluation of neural network models in traffic volume forecasting," *Math. Comput. Model.*, vol. 27, no. 9–11, pp. 293–310, May/Jun. 1998.

[24] R. Yasdi, "Prediction of road traffic using a neural network approach," *Neural Comput. Appl.*, vol. 8, no. 2, pp. 135–142, May 1999.

[25] B. Abdulhai, H. Porwal, and W. Recker, "Short-term freeway traffic flow prediction using genetically-optimized time-delay-based neural networks," presented at the 78th Annual Meeting Transportation Research Board, Washington, DC, USA, 1999.

[26] S. Ishak and C. Alecsandru, "Optimizing traffic prediction performance of neural networks under various topological, input and traffic condition settings," presented at the 82nd Annual Meeting Transportation Research Board, Washington, DC, USA, 2003.

[27] H. Dia, "An object-oriented neutral network approach to short-term traffic forecasting," *Eur. J. Oper. Res.*, vol. 131, no. 2, pp. 253–261, Jun. 2001.

[28] J. W. C. Van Lint, S. P. Hoogendoorn, and H. J. Van Zuylen, "Freeway travel time prediction with state-space neural networks: Modeling state-space dynamics with recurrent neural networks," presented at the 81st Annual Meeting Transportation Research Board, Washington, DC, USA, 2002.

[29] S. Sun, C. Zhang, and G. Yu, "A Bayesian network approach to traffic flow forecasting," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 1, pp. 124–132, Mar. 2006.

[30] S. Sun and M. Xu, "Variational inference for infinite mixtures of Gaussian processes with applications to traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 2, pp. 466–475, Jun. 2011.

[31] Y. Zhang and Z. Ye, "Short-term traffic flow forecasting using fuzzy logic system methods," *J. Intell. Transp. Syst.*, vol. 12, no. 3, pp. 102–112, 2008.

[32] W. Min and L. Wynter, "Real-time road traffic prediction with spatio-temporal correlations," *Transp. Res. Part C, Emerging Technol.*, vol. 19, no. 4, pp. 606–616, Aug. 2011.

[33] T. Pan, A. Sumalee, R. Zhong, and N. Indra-Payoong, "Short-term traffic state prediction based on temporal–spatial correlation," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 3, pp. 1242–1254, Sep. 2013.

[34] S. Sun, R. Huang, and Y. Gao, "Network-scale traffic modeling and forecasting with graphic lasso and neural networks," *J. Transp. Eng.*, vol. 138, no. 11, pp. 1358–1367, 2012.

[35] R. Huang and S. Sun, "Kernel regression with sparse metric learning," *J. Intell. Fuzzy Syst., Appl. Eng. Technol.*, vol. 24, no. 4, pp. 775–787, Jul. 2013.

[36] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Tree*. Belmont, CA, USA: Wadsworth, 1984, pp. 216–232.

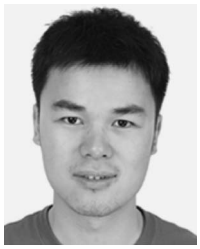
[37] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[38] T. Hastie, R. Tibshirani, and J. Friedman, *The Element of Statistical Learning*. New York, NY, USA: Springer-Verlag, 2009, pp. 411–615.

[39] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufman, 2011, pp. 199–200.

[40] T. Kim, D. Lovell, and J. Paracha, "A New Methodology to Estimate Capacity for Freeway Work Zones," presented at the 80th Annual Meeting Transportation Research Board, Washington, DC, USA, 2001.

[41] M. T. Hagan and M. Menhaj, "Training feed-forward networks with the Marquardt algorithm," *IEEE Trans. Neural Netw.*, vol. 5, no. 6, pp. 989–993, Nov. 1994.



Yi Hou received the B.S. degree in civil engineering from Southwest Jiaotong University, Chengdu, China, in 2004. He is currently working toward the Ph.D. degree in civil engineering at University of Missouri, Columbia, MO, USA.

His research interests include intelligent transportation system, traffic modeling and simulation, traffic safety, machine learning, and data mining.

Praveen Edara (M'14) received the Ph.D. degree in transportation systems from Virginia Polytechnic Institute and State University, Blacksburg, VA, USA.

He is an Associate Professor with University of Missouri, Columbia, MO, USA. His research interests include traffic modeling, simulation, and intelligent transportation system.

Carlos Sun received the B.S. degree in electrical engineering and the M.S. and Ph.D. degrees in civil engineering from University of California, Irvine, CA, USA, and the J.D. degree from University of Missouri, Columbia, MO, USA.

He is currently with the Department of Civil Engineering, University of Missouri. His research interests include intelligent transportation system, safety, work zones, and legal issues.