Travel Speed Forecasting by Means of Continuous Conditional Random Fields

Nemanja Djuric, Vladan Radosavljevic, Vladimir Coric, and Slobodan Vucetic

This paper explores the application of the recently proposed continuous conditional random fields (CCRF) to travel forecasting. CCRF is a flexible, probabilistic framework that can seamlessly incorporate multiple traffic predictors and exploit spatial and temporal correlations inherently present in traffic data. In addition to improving prediction accuracy, the probabilistic approach provides information about prediction uncertainty. Moreover, information about the relative importance of particular predictor and spatial–temporal correlations can be easily extracted from the model. CCRF is fault-tolerant and can provide predictions even when some observations are missing. Several CCRF models were applied to the problem of travel speed prediction in a range from 10 to 60 min ahead and evaluated on loop detector data from a 5.71-mi section of I-35W in Minneapolis, Minnesota. Several CCRF models, with increasing levels of complexity, are proposed to assess performance of the method better. When these CCRF models were compared with the linear regression model, they reduced the mean absolute error by around 4%. The results imply that modeling spatial and temporal neighborhoods in traffic data and combining various baseline predictors under the CCRF framework can be beneficial.

Reliable short-term travel forecasting is a crucial requirement in advanced traveler information systems. Accurate predictive modeling of traffic behavior leads to better vehicle guidance systems and trip planning, which can save both time and money for an ever-increasing number of travelers on freeways worldwide. A large variety of forecasting algorithms for traffic flow, speed, and travel time, which base their forecast on current traffic conditions and historical data, have been proposed and evaluated.

Baseline approaches that are typically used for benchmarking of more-advanced forecasting algorithms are the random-walk (RW) and historical models. The RW model predicts that future traffic conditions are identical to current ones. For extremely short-term forecasting (up to 10 min), the RW model is extremely accurate and hard to beat by more sophisticated models. The historical model predicts the future traffic state as the average of historical traffic under the same conditions (location, day of week, and time of day). The historical model is quite competitive in longer-term forecasting (a few hours ahead). Linear regression (*1, 2*) and its variants (*3*) have been popular because of their ease of implementation and ability to interpret and analyze the resulting model. Autoregressive integrated moving average (ARIMA) models, which encompass RW, random-trend models; auto-regressive models; and exponential weighted moving averages are linear time series models that have been quite popular thanks to their ability to exploit temporal dependence in prediction errors (*4, 5*). Linear models that exploit both spatial and temporal information that have been used in traffic forecasting include Kalman filters (*6*) and spatial–temporal ARIMA (*7*).

Two extensions of the linear models have been quite useful in traffic forecasting. By observing that a single set of parameters might not be suitable over varying traffic conditions, time-varying linear regression models have been employed with considerable success (*8–10*). As an alternative to a continuous change in model parameters, regime-switching models have been implemented (*10*) to model different traffic states separately, such as congested and free-flow regimes.

In addition to linear models, nonlinear ones have also been extensively studied in traffic-forecasting research. Neural networks (*11, 12*) and, more recently, support vector machines (*13*) have been used with some success; this success indicates that the nonlinearities in spatial–temporal traffic behavior can be exploited. The nonparametric approach based on the nearest neighbor algorithm, which is a generalization of the baseline historical method, has also been popular because of its ease of implementation and reasonable forecasting accuracy. For an extensive and quite good overview of traffic forecasting methods see Vlahogianni et al. (*14*).

This paper explores a recently proposed conditional random field (CRF) framework (*15*) that is extremely successful in modeling dependencies among random variables (such as spatial–temporal correlations) in forecasting problems. This novel method can be easily implemented and is quite flexible, and the goal of this research was to evaluate the method's applicability in traffic forecasting. Originally, CRFs were proposed for classification of sequential data (*15*), as an attractive alternative to hidden Markov models (*16*). Unlike those models, CRFs make no independence assumptions between input variables, and this avoidance of assumptions results in a more flexible modeling framework. Meanwhile, CRFs were applied in a number of fields, including computer vision (*17*) and computational biology (*18*). Recently, CRFs have been extended to solve regression problems. Continuous CRFs (CCRFs) were first described in Qin et al. (*19*) in the context of the problem of global document ranking. Following the Qin et al. work, an extension of CCRF that is applicable to spatial–temporal data was proposed by Radosavljevic et al. (*20*).

Because it can easily deal with real-valued, spatial–temporal data, CCRF naturally emerges as an extremely suitable model for travel forecasting. CCRF can combine various, possibly very different,

N. Djuric, 319 Wachman Hall; V. Radosavljevic, 323 Wachman Hall; V. Coric, 206 Wachman Hall; and S. Vucetic, 304 Wachman Hall, Department of Computer and Information Sciences, Temple University, 1805 North Broad Street, Philadelphia, PA 19122. Corresponding author: N. Djuric, nemanja.djuric@temple.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2263, Transportation Research Board of the National Academies, Washington, D.C., 2011, pp. 131–139.
DOI: 10.3141/2263-15

input data available in traffic forecast applications, such as sensor readings, historical data, information about weather conditions, incidents, and others, and can also model spatial–temporal correlations that inherently exist in traffic data. The employment of CCRF enables researchers to work in the probabilistic setting and to use the resulting model to calculate many statistical quantities, including the expected value of the output variables, their mode, and prediction uncertainty. One extremely important feature of CCRF is robustness to sensor failures, which are very common in traffic monitoring. While such failures pose a serious problem for many forecasting algorithms, CCRF can continue to provide predictions without any remedial action. The only effect of sensor failures on CCRF is increased forecasting uncertainty.

GENERAL CCRF MODEL

This section outlines a general CCRF model, which was introduced in Radosavljevic et al. (*20*). The next section illustrates how CCRFs can be implemented for a specific traffic forecasting problem. CCRFs provide a probabilistic framework for seamless incorporation of various aspects of a complex data dependence structure into a single model. Let us denote $\mathbf{x}$ as a vector of observations and $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ as an N-dimensional vector of real-valued output (or dependent) variables. A CCRF corresponds to the undirected graphical model illustrated in Figure 1, in which vertices representing the output variables $\mathbf{y}$ do not create cliques (defined as the subsets of fully connected vertices) larger than two. In this case, the conditional distribution $P(\mathbf{y}|\mathbf{x})$ for a CCRF can be represented in a convenient form as

$$P(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})} \exp\left(\sum_{i=1}^{N} A(\boldsymbol{\alpha}, y_i, \mathbf{x}) + \sum_{i \sim j} I(\boldsymbol{\beta}, y_i, y_j, \mathbf{x}) \right) \quad (1)$$

where

$A(\boldsymbol{\alpha}, y_i, \mathbf{x})$ = association potential with weights $\boldsymbol{\alpha}$,
$I(\boldsymbol{\beta}, y_i, y_j, \mathbf{x})$ = interaction potential with weights $\boldsymbol{\beta}$,
$i \sim j$ = that y_i and y_j are connected by an edge (neighbors), and
$Z(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ = normalization function defined as

$$Z(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int_{\mathbf{y}} \exp\left(\sum_{i=1}^{N} A(\boldsymbol{\alpha}, y_i, \mathbf{x}) + \sum_{i \sim j} I(\boldsymbol{\beta}, y_i, y_j, \mathbf{x}) \right) d\mathbf{y} \quad (2)$$

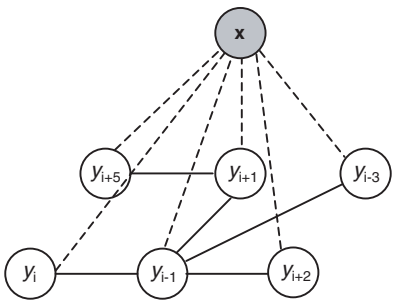

FIGURE 1 CCRF graphical structure (dashed line is association potential; solid line is interaction potential).

A large value of association potential A in a CCRF indicates that an output variable is closely related to the observations, while a large value of interaction potential I indicates stronger interaction between output variables. The design decisions in developing a CCRF model (Equation 1) include defining of the graph structure and deciding on the functional form of the association and interaction potentials. The next section illustrates how to develop a CCRF model for travel speed forecasting. As this paper will show, a graph similar to that in Figure 1 can be used as a basis for the spatial–temporal CCRF model for travel speed forecasting.

Given the CCRF model (Equation 1) and training set $D = \{(\mathbf{x}_t, \mathbf{y}_t), t = 1, 2, \ldots, T\}$, the task is to find weights $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ such that conditional log likelihood of the training data $L(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is maximized:

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{t=1}^{T} \log P(\mathbf{y}_t \mid \mathbf{x}_t)$$
$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \arg\max_{\alpha, \beta} (L(\boldsymbol{\alpha}, \boldsymbol{\beta})) \quad (3)$$

To avoid overfitting, regularization terms $\frac{1}{2}\boldsymbol{\alpha}^2$ and $\frac{1}{2}\boldsymbol{\beta}^2$ are usually added to the log likelihood $L(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in Equation 3. The optimization problem (Equation 3) is typically solved by using the gradient ascent. Once the CCRF model is trained, one can evaluate the conditional probability for a new observation $\mathbf{x}_{new}$. Quite often, users are interested only in the most likely value of y_{new} given $\mathbf{x}_{new}$:

$$\hat{y}_{new} = \arg\max_{y_{new}} (P(y_{new} | \mathbf{x}_{new})) \quad (4)$$

In general, both learning (Equation 3) and inference (Equation 4) can be extremely difficult to solve. However, if one pays attention during design of the CCRF model, both learning and inference can be accomplished in a computationally efficient manner. The remainder of this section gives some general guidelines for designing computationally tractable CCRF models.

The first step toward tractable CCRF is that both association and interaction potentials are linear functions in the model weights $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$:

$$A(\boldsymbol{\alpha}, y_i, \mathbf{x}) = \sum_{m=1}^{M} \alpha_{mi} f_{mi}(y_i, \mathbf{x})$$
$$I(\boldsymbol{\beta}, y_i, y_j, \mathbf{x}) = \sum_{l=1}^{L} \beta_l g_l(y_i, y_j, \mathbf{x}) \quad (5)$$

where

f_{mi} and g_l = feature functions, with weights α_{mi} and β_l, respectively;
M = number of prediction models (baseline predictors); and
L = number of interaction definitions included in the model (temporal, spatial, or any other).

A large number of feature functions can be introduced because their actual relevance will be automatically determined during training. More relevant features will be given bigger weights, whereas the irrelevant ones will get smaller weights and their influence on the CCRF model will be smaller.

The second requirement that leads to a tractable and easy-to-interpret CCRF model is that the feature functions are quadratic

functions of output variables $\mathbf{y}$. It can be shown that in this case the conditional distribution $P(\mathbf{y}|\mathbf{x})$ corresponds to the multivariate Gaussian distribution (*20*). With this in mind, this paper considers a class of feature functions for association potential defined as

$$f_{mi}(y_i, \mathbf{x}) = -\delta_{mi}(\mathbf{x})(y_i - \theta_{mi}(\mathbf{x}))^2 \tag{6}$$

where θ_{mi} are the predictor functions and δ_{mi} are the indicator functions having values 0 or 1. Predictor function θ_{mi} can be any baseline predictor of output variable y_i. Therefore, definition (Equation 6) allows the use of M different predictors for each output variable. On the basis of Equation 6, predictions θ_{mi} that are the most accurate will result in the highest values of the feature function. The indicator functions are optional choices and can be used to improve the expressiveness of the resulting CCRF model. The next section will predict how the predictor and indicator functions can be selected in the case of travel speed forecasting. Feature functions for the interaction potential that will be considered are of the form

$$g_l(y_i, y_j) = -w^l(y_i, y_j)(y_i - y_j)^2 \tag{7}$$

where $w^l(y_i, y_j)$ is a nonnegative weight function representing a measure of correlation between the ith and jth outputs under the lth neighborhood definition. For instance, in traffic forecasting, $w^l(y_i, y_j)$ could measure temporal or spatial proximity of the ith and jth outputs. If two outputs are neighbors, their difference will have an influence on the value of the interaction potential. In general, interaction potential (Equation 5) can depend on observations $\mathbf{x}$. Therefore, Equation 7 can be extended by adding dependence on $\mathbf{x}$. However, the current case study did not use the observation variables $\mathbf{x}$ in the interaction potential and for simplicity of presentation omitted them from Equation 7.

After Equations 5 through 7 are combined with Equation 1, the resulting CCRF model can be expressed as

$$P(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})} \exp\left(\begin{array}{l} -\sum_{i=1}^{N}\sum_{m=1}^{M} \alpha_{mi}\delta_{mi}(\mathbf{x})(y_i - \theta_{mi}(\mathbf{x}))^2 \\ -\sum_{i\sim j}\sum_{l=1}^{L} \beta_l w^l(y_i, y_j)(y_i - y_j)^2 \end{array} \right) \tag{8}$$

As shown in Radosavljevic et al. (*20*), thus defined conditional distribution $P(\mathbf{y} \mid \mathbf{x})$ now corresponds to multivariate Gaussian distribution:

$$P(\mathbf{y} \mid \mathbf{x}) = \frac{1}{(2\pi)^{N/2}\left|\boldsymbol{\Sigma}(\mathbf{x})\right|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}(\mathbf{x}))^T \boldsymbol{\Sigma}(\mathbf{x})^{-1}(\mathbf{y} - \boldsymbol{\mu}(\mathbf{x})) \right) \tag{9}$$

with mean $\boldsymbol{\mu}(\mathbf{x})$ and covariance matrix $\boldsymbol{\Sigma}(\mathbf{x})$ that will be defined in the following. Both the mean and covariance matrix are not constant, and they depend on $\mathbf{x}$. The remainder of this section, for the simplicity of notation, will not explicitly write this dependency on $\mathbf{x}$. N-dimensional vector $\mathbf{b}$ is defined with the following elements:

$$b_i = 2\sum_{m=1}^{M} \alpha_{mi}\delta_{mi}(\mathbf{x})\theta_{mi}(\mathbf{x}) \tag{10}$$

and $N \times N$ matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ with the elements

$$\mathbf{Q}_{1ij} = \begin{cases} \sum_{m=1}^{M} \alpha_{mi}\delta_{mi}(\mathbf{x}) & i = j \\ 0 & \text{otherwise} \end{cases}$$
$$\mathbf{Q}_{2ij} = \begin{cases} \sum_{j=1}^{N}\sum_{l=1}^{L} \beta_l w^l(y_i, y_j) & i = j \\ -\sum_{l=1}^{L} \beta_l w^l(y_i, y_j) & i \neq j \end{cases} \tag{11}$$

One can observe that $\mathbf{Q}_1$ is a diagonal matrix and $\mathbf{Q}_2$ is a sparse matrix that reflects the dependency structure between output variables. Now, as shown in Radosavljevic et al. (*20*), the inverse of the covariance matrix can be calculated as

$$\boldsymbol{\Sigma}^{-1} = 2(\mathbf{Q}_1 + \mathbf{Q}_2) \tag{12}$$

and the mean of conditional distribution as

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{b} \tag{13}$$

Now one can revisit questions of training of the CCRF model and of using the trained model in forecasting. Weights $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ must remain positive during training, and this requirement constrains the optimization problem. If that were not the case, feature functions would lose the property that their values are bigger when the prediction models are more accurate (in the case of weights $\boldsymbol{\alpha}$) or when the outputs are more temporally or spatially related (in the case of weights $\boldsymbol{\beta}$). Therefore, in a manner similar to that in Qin et al. (*19*) and Radosavljevic et al. (*20*), the log likelihood will be maximized with respect to the logarithms of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The resulting optimization problem is now unconstrained, and the gradient ascent can simply be used for the model to learn the weights as

$$\log\boldsymbol{\alpha}^{\text{new}} = \log\boldsymbol{\alpha}^{\text{old}} + \eta\frac{\partial L}{\partial \log\boldsymbol{\alpha}^{\text{old}}}$$
$$\log\boldsymbol{\beta}^{\text{new}} = \log\boldsymbol{\beta}^{\text{old}} + \eta\frac{\partial L}{\partial \log\boldsymbol{\beta}^{\text{old}}} \tag{14}$$

where η is the learning rate.

For the forecasting, given the observation $\mathbf{x}$**,** the most likely prediction for vector $\mathbf{y}$ equals the mean of the conditional distribution:

$$\hat{\mathbf{y}} = \boldsymbol{\mu}(\mathbf{x}) = \boldsymbol{\Sigma}\mathbf{b} \tag{15}$$

The 95% confidence intervals are estimated from mean $\hat{\mathbf{y}}$ and covariance matrix $\boldsymbol{\Sigma}$ and with consideration of the following formula that stems from the Gaussian framework:

$$P(\hat{\mathbf{y}} - 1.96 \cdot \text{diag}(\boldsymbol{\Sigma}) \leq \mathbf{y} \leq \hat{\mathbf{y}} + 1.96 \cdot \text{diag}(\boldsymbol{\Sigma})) = 0.95 \tag{16}$$

where $\text{diag}(\boldsymbol{\Sigma})$ denotes the main diagonal of $\boldsymbol{\Sigma}$. This equation means that, with a probability of 95%, the random variable $\mathbf{y}$ will fall in a range defined by $(\hat{\mathbf{y}} \pm 1.96 \cdot \text{diag}(\boldsymbol{\Sigma}))$.

A special case of Equation 15 exists. As Equations 10 and 11 show, vector **b** and matrix $\mathbf{Q}_1$ do not depend on the interaction potential. Therefore, if the interaction potential is not used (output variables are assumed to be conditionally independent), the prediction for y_i simply entails a weighted average of the prediction functions and can be expressed as

$$y_i = \frac{\sum_{m=1}^{M} \alpha_{mi}\delta_{mi}(\mathbf{x})\theta_{mi}(\mathbf{x})}{\sum_{m=1}^{M} \alpha_{mi}\delta_{mi}(\mathbf{x})} \tag{17}$$

Finally, a discussion of the computational costs of model training and inference will be helpful. If the size of the training set is T and training stops after K iterations, it takes $O(T \cdot K \cdot N^3)$ time to train the model because, during gradient ascent, the inverse of a block-diagonal covariance matrix $\boldsymbol{\Sigma}$ must be found. However, this is the worst-case performance because matrix $\boldsymbol{\Sigma}$ is usually extremely sparse and training time can be decreased by several orders of magnitude. Training can be done offline. However, during inference, which is done online, the product of covariance matrix $\boldsymbol{\Sigma}$ and vector **b**, which takes $O(N^2)$ time, simply needs to be found. If again sparsity of matrix $\boldsymbol{\Sigma}$ is again exploited, time cost can drop as low as $O(N)$.

CCRF MODELS FOR TRAVEL SPEED FORECASTING

CCRF and Travel Speed Forecasting

CCRF is a general framework that allows significant modeling flexibility. The following case study is aimed at illustrating the potential of CCRF in traffic forecasting. The specific problem being studied is travel speed forecasting. In the considered scenario, the objective is to predict, at any given time t, the travel speed across S adjacent sensor stations of the freeway over H forecasting horizons. For time t, the input variables $\mathbf{x}_t$ are defined as a set of traffic observations available at time t, including current and historical observations and their derivatives. Output variables $\mathbf{y}_t$ (travel speeds) form a vector of size $N = S \cdot H$. For the convenience of notation, the remainder of this section will represent $\mathbf{y}_t$ as a matrix of size $S \times H$ with elements $\{y_{s,t+h}, s = 1, 2, \ldots, S; h = 1, 2, \ldots, H\}$, where $y_{s,t+h}$ is travel speed on station s at time $t + h$, S is the number of stations, and H is the number of time horizons.

CCRF Model Variants

This section proposes several CCRF models with increasing levels of complexity. This increasing complexity will help readers understand the flexibility and strength of the CCRF framework.

Case 1. Basic Model

The basic model uses only two baseline predictors while indicator functions and interaction potential are omitted. For simplicity of presentation, only the simplest predictors, RW and historical predictors, are used. A more complex predictor, such as linear regression, ARIMA, or neural networks, could be used instead without any impact on computational cost of training CCRF and without its being used for forecasting. The RW predictor simply uses the current speed on the station to predict future speed for every horizon, and the historical-median predictor (HM) uses the median historical speed $h_{s,t+h}$ for the given station s and time $t + h$. Here, median is used instead of mean because median is more robust to outliers and achieves better accuracies. More formally, two feature functions f_1 and f_2 are introduced as

$$\begin{aligned} f_1(y_{s,t+h}, \mathbf{x}_t) &= -\left(y_{s,t+h} - \mathrm{RW}_{s,t+h}(\mathbf{x}_t)\right)^2 = -\left(y_{s,t+h} - y_{s,t}\right)^2 \\ f_2(y_{s,t+h}, \mathbf{x}_t) &= -\left(y_{s,t+h} - \mathrm{HM}_{s,t+h}(\mathbf{x}_t)\right)^2 = -\left(y_{s,t+h} - h_{s,t+h}\right)^2 \end{aligned} \tag{18}$$

where $\mathrm{RW}_{s,t+h}(\mathbf{x}_t) = y_{s,t}$ and $\mathrm{HM}_{s,t+h}(\mathbf{x}_t) = h_{s,t+h}$ are predicted travel speeds for station s and time $t + h$ given by the current time and historical median speed, respectively. For this model, the resulting conditional distribution is

$$P(\mathbf{y}_t|\mathbf{x}_t) = \frac{1}{Z}\exp\begin{pmatrix} -\sum_{s=1}^{S}\sum_{h=1}^{H} \alpha_{1sh}\left(y_{s,t+h} - y_{s,t}\right)^2 \\ -\sum_{s=1}^{S}\sum_{h=1}^{H} \alpha_{2sh}\left(y_{s,t+h} - h_{s,t+h}\right)^2 \end{pmatrix} \tag{19}$$

where α_{1sh} and α_{2sh} are weights that measure goodness of the RW and HM predictors on station s at time horizon h, respectively. It is expected that in the trained CCRF, the accurate baseline predictors would have large values of the associated weights $\boldsymbol{\alpha}$ and large influence on the forecasting of speed. Less-accurate predictors would have low weight and marginal influence on forecasting. The total number of weights in this model is $2S \cdot H$. If the number of training iterations is constant, the time needed to train this model is only $O(T \cdot H \cdot S)$ because the covariance matrix $\boldsymbol{\Sigma}$ is diagonal and its inversion takes linear time. Forecasting of all $H \cdot S$ speeds is extremely fast and takes $O(H \cdot S)$ time.

Case 2. Simple Model

For the simple model, two new baseline predictors are introduced while indicator functions and interaction potential are still omitted. The two new feature functions, f_3 and f_4, are

$$\begin{aligned} f_3(y_{s,t+h}, \mathbf{x}_t) &= -\left(y_{s,t+h} - y_{s-1,t}\right)^2 \\ f_4(y_{s,t+h}, \mathbf{x}_t) &= -\left(y_{s,t+h} - y_{s+1,t}\right)^2 \end{aligned} \tag{20}$$

These functions account for the current speed on spatial neighbors of station s at time t. Because the future state of traffic on the current station inevitably depends on the state of traffic on neighboring stations, this model is expected to capture the dynamics of traffic more faithfully and, in effect, achieve higher accuracy. The conditional distribution for this model is quite similar to that for the previous case, the only difference being that new weights, α_{3sh} and α_{4sh}, are introduced to measure goodness of the two new predictors. The total number of weights in this model is $4 \cdot H \cdot S$.

Case 3. Regime-Switching Model

Traffic behaves significantly differently during free flow than during the congested state. Therefore, extension to the simple model is

to build separate models for the two regimes (free-flow regime and congested state). To achieve this extension, the use of the indicator functions for the traffic regime is proposed. The model indicates the free-flow regime if the current speed is greater than 30 mph and the peak hour regime otherwise. The threshold of 30 mph was taken after preliminary studies showed that it can be an appropriate choice. Now, the conditional distribution can be written as

$$P(\mathbf{y}_t|\mathbf{x}_t)=\frac{1}{Z}\exp\Bigg(-\sum_{j=1}^{2}\sum_{s=1}^{S}\sum_{h=1}^{H}\alpha_{1jsh}\delta_j(y_{s,t})(y_{s,t+h}-y_{s,t})^2 -\sum_{j=1}^{2}\sum_{s=1}^{S}\sum_{h=1}^{H}\alpha_{2jsh}\delta_j(y_{s,t})(y_{s,t+h}-h_{s,t+h})^2 -\sum_{j=1}^{2}\sum_{s=1}^{S}\sum_{h=1}^{H}\alpha_{3jsh}\delta_j(y_{s,t})(y_{s,t+h}-y_{s-1,t})^2 -\sum_{j=1}^{2}\sum_{s=1}^{S}\sum_{h=1}^{H}\alpha_{4jsh}\delta_j(y_{s,t})(y_{s,t+h}-y_{s+1,t})^2\Bigg) \quad (21)$$

where α_{mjsh} weights measure goodness of the mth predictors on station s at forecasting horizon h when the traffic is in the jth regime. The indicator functions are defined as $\delta_1(z) = 1$, if $z \le 30$ mph and $\delta_1(z) = 0$ otherwise, and $\delta_2(z) = 1$, if $z > 30$ mph and $\delta_2(z) = 0$, otherwise. In this way, the use of these functions allows assignment of different weights for free-flow and congested regimes. CCRF allows the use of more-sophisticated detectors of congested regimes to define the indicator functions, but this idea was not pursued further here. The total number of weights in the regime-switching model is $8 \cdot S \cdot H$.

Case 4. Correlations Model

To model the interaction potential as defined in Equation 7, agreement is needed on the dependence relations between output variables. In the forecasting of travel speed, it is useful to give two definitions of "neighborhood between outputs," temporal and spatial. In Figure 2, temporal neighbors are connected with a thin solid line while spatial neighbors are connected with a thick solid line. Formally put, output variables $y_{s1,h1}$ and $y_{s2,h2}$ are said to be spatial neighbors if $h_1 = h_2$ and $|s_1 - s_2| = 1$. Conversely, outputs $y_{s1,h1}$ and $y_{s2,h2}$ are said to be temporal neighbors if $s_1 = s_2$ and $|h_1 - h_2| = 1$. Two weight functions w^l from Equation 7 are defined as

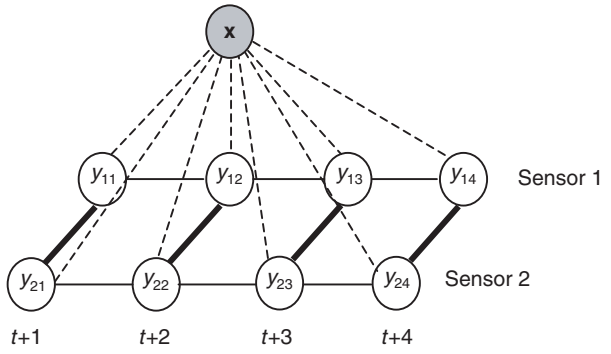

FIGURE 2 CCRF graphical model suitable for traffic forecast (dashed line is association potential; solid line is interaction potential).

$$w^1(y_{s_1,h_1},y_{s_2,h_2})=\begin{cases}1 & \text{if } y_{s_1,h_1} \text{ and } y_{s_2,h_2} \text{ are temporal neighbors}\\ 0 & \text{otherwise}\end{cases}$$
$$w^2(y_{s_1,h_1},y_{s_2,h_2})=\begin{cases}1 & \text{if } y_{s_1,h_1} \text{ and } y_{s_2,h_2} \text{ are spatial neighbors}\\ 0 & \text{otherwise}\end{cases} \quad (22)$$

In this work, weight function w^l is defined as a binary indicator function, although, in general, it can be any nonnegative, real-valued function. Finally, the conditional distribution could be written as

$$P(\mathbf{y}_t|\mathbf{x}_t)=\frac{1}{Z}\exp\Bigg(-\sum_{j=1}^{2}\sum_{s=1}^{S}\sum_{h=1}^{H}\alpha_{1jsh}\delta_j(y_{s,t})(y_{s,t+h}-y_{s,t})^2 -\sum_{j=1}^{2}\sum_{s=1}^{S}\sum_{h=1}^{H}\alpha_{2jsh}\delta_j(y_{s,t})(y_{s,t+h}-h_{s,t+h})^2 -\sum_{j=1}^{2}\sum_{s=1}^{S}\sum_{h=1}^{H}\alpha_{3jsh}\delta_j(y_{s,t})(y_{s,t+h}-y_{s-1,t})^2 -\sum_{j=1}^{2}\sum_{s=1}^{S}\sum_{h=1}^{H}\alpha_{4jsh}\delta_j(y_{s,t})(y_{s,t+h}-y_{s+1,t})^2 -\sum_{s=1}^{S}\sum_{h=1}^{H-1}\beta_{1sh}(y_{s,t+h}-y_{s,t+h+1})^2 -\sum_{s=1}^{S-1}\sum_{h=1}^{H}\beta_{2sh}(y_{s,t+h}-y_{s+1,t+h})^2\Bigg) \quad (23)$$

where β_{1sh} weight measures the strength of temporal interaction on the sth station between the hth and the hth + 1 forecasting horizons, and β_{2sh} weight measures the strength of spatial interaction between sth and sth + 1 stations, at the hth forecasting horizon. If the spatial or temporal interaction is strong, the corresponding weight will be assigned a higher value. The total number of weights in the final model is $10 \cdot S \cdot H - H - S$, where $8 \cdot S \cdot H$ are used to model association potential, while $(S - 1) \cdot H$ of spatial weights and $S \cdot (H - 1)$ of temporal weights are used to model interaction potential.

The correlations model is the most complex one considered in this paper, which has demonstrated how the model can be easily extended, however deemed appropriate, by using different feature and indicator functions. Possible avenues for extensions are (*a*) to define weight functions w^l as real-valued functions that would, for instance, depend on the length of the road sections or on the size of the time step or (*b*) to include additional baseline predictors (neural networks, linear regression, ARIMA). Alternative neighborhood definitions, in addition to the two defined in this paper, can also be included. Extensions are introduced seamlessly without sacrificing training time, and the actual significance of every new feature function will automatically be determined during training.

EXPERIMENTAL RESULTS

The proposed CCRF models were tested on the loop data from a 5.71-mi section of I-35W through Minneapolis, Minnesota. The data were collected from April 1 to July 1, 2003, between 2 and 7 p.m. This data set was selected because it contained a congested regime that occurred almost regularly during afternoon hours. When weekends, holidays, and days with a large fraction of missing or corrupted data were removed, 43 days remained in the data set, which

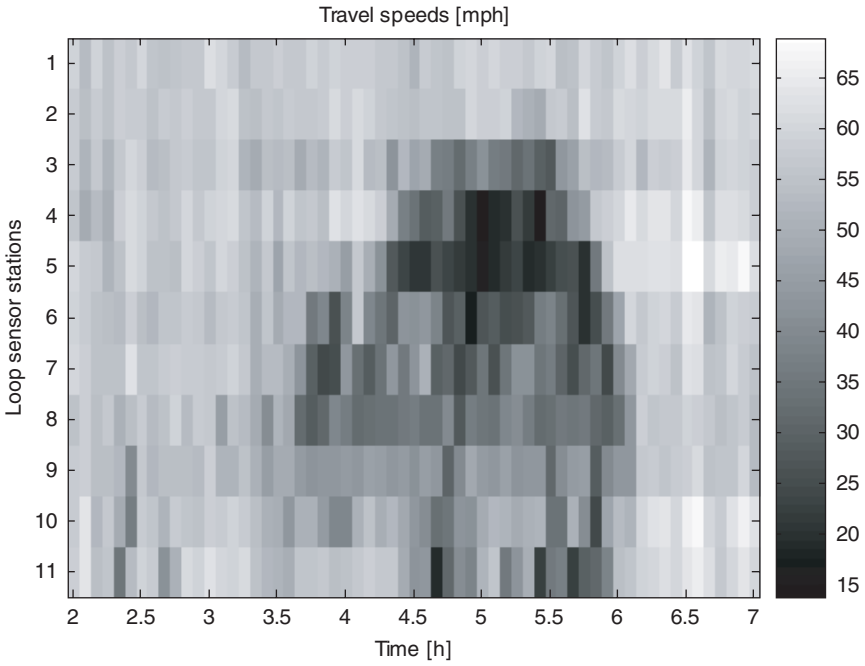

FIGURE 3 Typical day with sizeable congestion period (April 7, 2003).

consisted of travel speeds from 11 ($S = 11$) consecutive loop sensors, aggregated over 5-min intervals. The traffic flowed from Station 1 to Station 11. Figure 3 shows travel speeds on this road segment between 2 and 7 p.m. on April 7, 2003. Travel speeds were derived from the 30-s loop sensor data that measured volume (the number of cars that passed the sensor during a 30-s interval) and occupancy (the time that the sensor was occupied during a 30-s interval). For every station, a method similar to that in Coifman (*21*) was used to estimate the average vehicle length as

$$\text{length} = 60 \text{ mph} \times \text{median}\left(\frac{\text{occupancy}}{\text{volume}}\right) \tag{24}$$

where median was calculated during the free-flow traffic conditions and the constant free-flow speed of 60 mph was assumed. Also assumed was that the average vehicle length does not change during different periods of the day. Then, the travel speed for the given sensor was calculated as

$$\text{speed} = \text{length} \times \frac{\text{volume}}{\text{occupancy}} \tag{25}$$

The task was to predict, in 5-min increments, the speed on $S = 11$ stations over six forecasting horizons ($H = 6$), namely, 10, 20, 30, 40, 50, and 60 min ahead. The size of the data set T was 43×49 (forty-nine 5-min periods covering between 2 and 7 p.m. during 43 days) = 2,107. Three cross-validation experiments were conducted and mean absolute error (MAE) was reported. The data set was divided into three subsets that contained data from 15, 15, and 13 days. Then, the CCRF models were trained on two of the subsets and tested on the remaining one. This procedure was repeated three times such that each subset was used once for testing. The reported accuracy is the average over the three repetitions.

Table 1 summarizes the MAE of various models over six time horizons as well as the overall MAE. For comparison with the four proposed CCRF models, report accuracies of the following were also reported:

- Linear regression (LR) models. For LR, two data-generating processes were assumed: $y_{s,t+h} = \alpha_1 y_{s,t} + \alpha_2 h_{s,t+h} + \epsilon_{s,t+h}$ (with two baselines equivalent to CCRF Model 1) and $y_{s,t+h} = \alpha_1 y_{s,t} + \alpha_2 h_{s,t+h} + \alpha_3 y_{s-1,t} + \alpha_4 y_{s+1,t} + \epsilon_{s,t+h}$ (with four baselines, equivalent to CCRF Model 2). For each LR model, 6×11 LR predictors were modeled, one for each station and for each forecasting horizon. The ordinary least squares algorithm was used to learn weights α_i.
- Baseline predictors. Four baseline predictors were used: (*a*) RW that uses current speed at the station, (*b*) historical speed at the station, (*c*) current speed at the station upstream of the sensor, and (*d*) current speed at the station downstream of the station. The first two baseline predictors are used in CCRF Model 1, while all four are used in the remaining CCRF models reported in Table 1.

As Table 1 shows, the RW predictor is better than the historical predictor over all six forecasting horizons, and the difference between them gradually decreases with the horizon length. RW based on the neighboring stations was slightly worse than the RW baseline predictor. Both LR models are significantly more accurate than either of the baseline predictors, indicating the benefits of combining current and historical speeds in forecasting. As expected, the CCRF Model 1 had similar accuracy to LR Model 1. However, the accuracy of CCRF Model 2, which uses current speeds of the neighboring stations, is higher than the accuracy of related LR Model 2. The use of regime indicator functions in CCRF Model 3 was helpful and led to an additional accuracy improvement. Finally, CCRF Model 4, which exploits spatial and temporal correlations, is the most accurate forecasting model. The improvement over CCRF Model 3 is relatively small.

TABLE 1 Prediction MAE Errors of Various Predictors Aggregated Across Time Horizons

	Horizon (min)						
	+10	+20	+30	+40	+50	+60	Total
Linear Regression Models							
LR Model 1 (two baselines)	6.096	7.634	8.755	9.821	10.605	11.136	9.008
LR Model 2 (four baselines)	5.961	7.547	8.724	9.807	10.625	11.154	8.969
Baseline Predictors Incorporated into CCRF Models							
Random walk	6.130	7.667	8.789	10.033	11.072	11.947	9.273
Historical median	13.090	13.183	13.153	12.994	12.737	12.419	12.931
Current speed of upstream sensor	7.568	8.746	9.789	10.980	11.995	12.849	10.321
Current speed of downstream sensor	7.311	8.627	9.543	10.582	11.505	12.277	9.974
CCRF Models, in Increasing Level of Complexity							
CCRF Model 1 (two baselines)	6.198	7.703	8.752	9.778	10.597	11.239	9.004
CCRF Model 2 (four baselines)	5.929	7.325	8.333	9.384	10.290	11.061	8.720
CCRF Model 3 (regime switching)	5.922	7.327	8.329	9.363	10.214	10.891	8.675
CCRF Model 4 (with correlations)	5.920	7.308	8.314	9.352	10.213	10.905	8.669

Table 2 reports the values of $\boldsymbol{\alpha}$ weights for Sensor Station 5, for CCRF Model 4. These weights are extremely informative because their absolute values indicate confidence of CCRF model in a given baseline predictor, and their relative values indicate the baseline predictors that the CCRF model trusts most. As Table 2 shows, RW (current) weights gradually decrease with the prediction horizon, indicating increased forecasting uncertainty. The weights of the historical predictor are smaller than those for the RW, but they remain relatively steady. This weight difference means that CCRF puts larger trust in the current speed than in the historical speed, especially on shorter forecasting horizons. Interesting differences also exist between the free-flow and the congested CCRF weights. In the congested (peak hour) regime in Table 2, current speed from the downstream station is given greater weight than is that from the upstream sensor. The reason for this greater weight is probably because the downstream station carries useful information about the ending of the congested regime. The relative importance of historical speed is much larger during the free-flow regime. Also recorded were the values of $\boldsymbol{\beta}$ weights modeling temporal and spatial correlation in predicted speeds. Values of temporal interaction weights β_{1sh} are around 0.01, which is somewhat lower than reported α_{jm5h} weights. This difference implies that temporal interactions between outputs are nearly as important as predictions of baseline predictors. In contrast, spatial interaction weights β_{2sh} were < 0.001, indicating that spatial dependency between traffic speeds is relatively weak. This statement can be explained by examination of Figure 3, in which large differences between travel speeds at neighboring sensors can be observed quite often.

Figures 4 and 5 correspond to CCRF Model 4. One of the advantages of CCRF is that, because the conditional probability is Gaussian, one can easily extract confidence intervals for the predictions. Figure 4 shows predictions with 95% confidence intervals for 10-min forecasting horizon for June 25, 2003, at Sensor Station 5. The CCRF model has a much-lower forecasting uncertainty during the free-flow regime than predictions made during the peak hour. This positive result indicates that CCRF is quite successful in estimating the

TABLE 2 Weights (α) for Four Baseline Methods Across Six Values of Time Horizon for Station 5

	Horizon (min)					
α Weight	+10	+20	+30	+40	+50	+60
$j = 1$ (rush hour)						
Current (α_{115h})	0.02309	0.01553	0.01270	0.00995	0.00863	0.00746
History (α_{215h})	0.00243	0.00264	0.00289	0.00295	0.00299	0.00302
Current – (α_{315h})	0.00815	0.00741	0.00672	0.00566	0.00507	0.00477
Current + (α_{415h})	0.01865	0.01360	0.01162	0.00993	0.00912	0.00783
$j = 2$ (free flow)						
Current (α_{125h})	0.02244	0.01052	0.00698	0.00537	0.00459	0.00408
History (α_{225h})	0.00682	0.00597	0.00542	0.00514	0.00494	0.00492
Current – (α_{325h})	0.01657	0.00853	0.00596	0.00470	0.00404	0.00362
Current + (α_{425h})	0.01420	0.01153	0.00877	0.00689	0.00585	0.00546

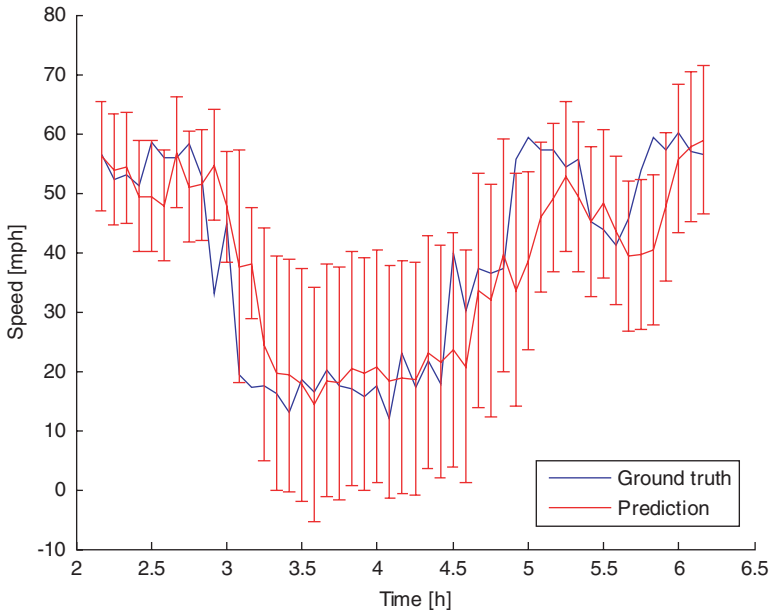

FIGURE 4 Confidence intervals of predictions (horizon = 10 min).

forecasting uncertainty. The error bars show that the 95% confidence interval was missed only a few times, most notably during the regime-change periods. This result is expected because the indicator functions in Radosavljevic et al. (*20*) are reactive in the sense that they are activated only when the regime change is already observed. This could be resolved by development and use of predictive indicator functions that would attempt to predict a free-flow–congested regime. As the forecasting horizon increases, the reported uncertainty by CCRF also increases (although this is not illustrated).

As noted early in the paper, another powerful feature of CCRF models is that they can make predictions even when some baseline predictors stop functioning (in other words, when some sensors malfunction and stop feeding information into the model). Figure 5 shows predictions for a 10-min forecasting horizon for June 25, 2003, at Sensor Station 5, where the second and the third baseline predictors (feature functions f_2 and f_3) stopped working from 2:10 to 2:40 p.m. Even when this malfunction occurs, the model outputs the speed predictions and reports the increased forecasting uncertainty, reflecting the partial loss of information.

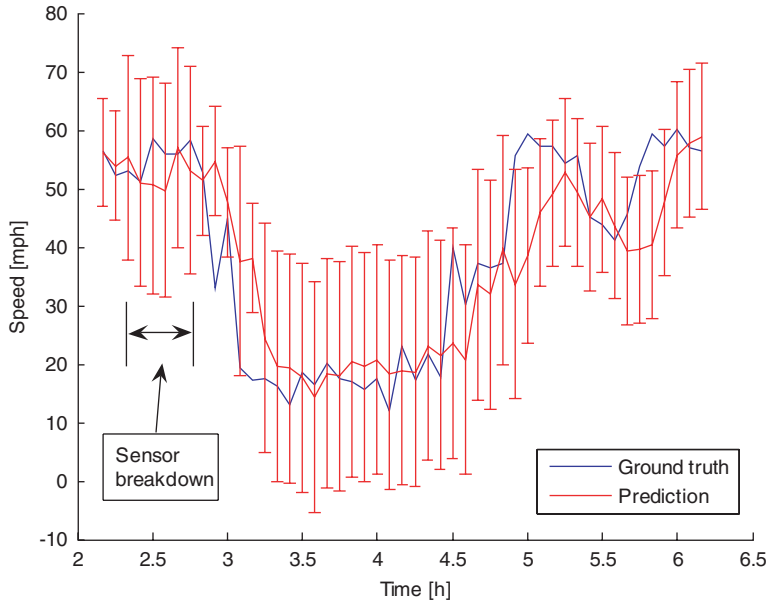

FIGURE 5 Confidence intervals of predictions (horizon = 10 min) when sensor breakdown occurred from 2:10 to 2:40 p.m.

CONCLUSION

This paper evaluated the recently proposed CCRF on the problem of travel speed forecasting. This extremely powerful probabilistic model can easily incorporate multiple baseline traffic predictors, model traffic regimes, and exploit temporal and spatial correlations within a single model. The resulting model is relatively simple and computationally efficient if the tractability requirements are followed. Several interesting and powerful properties of the CCRF model have been demonstrated. First, as input to the model, as many baseline predictors as desired can be included; the model will automatically detect the usefulness and reliability of each predictor and reflect this knowledge by assigning appropriate weights. Similarly, there is wide flexibility in modeling the spatial and temporal correlation between the output variables. In addition, the CCRF model can easily provide forecasting uncertainty estimates. Finally, the CCRF model is extremely robust to sensor failures, which is a common problem for traffic monitoring systems.

The four models used to illustrate application of the CCRF to traffic speed forecasting are probably not sufficient for predictions during special events (snowstorms, sports events, incidents). The reason for this possible insufficiency is that the baseline predictors used for feature functions are simple (current speeds and historical speeds) and the same holds for the indicator function (a simple detector of a free-flow or congested regime). Preliminary results indicate that, by using more sophisticated baseline predictors and indicator functions, the accuracy of the CCRF during nonrecurrent events could be improved. A topic of future research is to study enhancing CCRF models to improve accuracy during congested regimes, transitions between free-flow and congested regimes, and special events. It is evident that the proposed method can be applied to other traffic forecasting problems that go beyond speed prediction.

ACKNOWLEDGMENTS

This work was funded in part by the National Science Foundation. The authors are grateful to Taek Kwon of the University of Minnesota, Duluth, for providing access to the Minnesota Department of Transportation Traffic Management Center data.

REFERENCES

1. Kwon, J., and K. Petty. Travel Time Prediction Algorithm Scalable to Freeway Networks with Many Nodes with Arbitrary Travel Routes. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1935,* Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 147–153.
2. Rice, J., and E. van Zwet. A Simple and Effective Method for Predicting Travel Times on Freeways. *IEEE Transactions on Intelligent Transport Systems,* 2004, pp. 200–207.
3. Huang, L., and M. Barth. A Novel Loglinear Model for Freeway Travel Time Prediction. *Proc., 11th International IEEE Conference on Intelligent Transportation Systems,* Beijing, China, Oct. 12–15, 2008.
4. Van der Voort, M., M. Dougherty, and S. Watson. Combining KOHONEN Maps with ARIMA Time Series Models to Forecast Traffic Flow. *Transportation Research Part C,* Vol. 4, 1996, pp. 307–318.
5. Williams, B. M. Multivariate Vehicular Traffic Flow Prediction: Evaluation of ARIMAX Modeling. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1776,* TRB, National Research Council, Washington, D.C., 2001, pp. 194–200.
6. Chien, S. I. J., and C. M. Kuchipudi. Dynamic Travel Time Prediction with Real-Time and Historical Data. Presented at 81st Annual Meeting of the Transportation Research Board, Washington, D.C., 2002.
7. Williams, B. M., P. K. Durvasula, and D. E. Brown. Urban Freeway Traffic Flow Prediction: Application of Seasonal Autoregressive Integrated Moving Average and Exponential Smoothing Models. In *Transportation Research Record 1644,* TRB, National Research Council, Washington, D.C., 1998, pp. 132–141.
8. Hastie, T. J., and R. J. Tibshirani. Varying-Coefficient Models. *Journal of the Royal Statistical Society, Series B,* Vol. 55, 1993, pp. 757–796.
9. Zhang, X., and J. Rice. Short-Term Travel Time Prediction Using a Time-Varying Coefficient Linear Model. *Transportation Research Part C,* Vol. 11, 2001, pp. 187–210.
10. Cetin, M., and G. Comert. Short-Term Traffic Flow Prediction with Regime Switching Models. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1965,* Transportation Research Board of the National Academies, Washington, D.C., 2006, pp. 23–31.
11. Park, D., and L. R. Rilett. Forecasting Multiple-Period Freeway Link Travel Times Using Modular Neural Networks. In *Transportation Research Record 1617,* TRB, National Research Council, Washington, D.C., 1998, pp. 163–170.
12. van Lint, J. W. C., 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. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1811,* Transportation Research Board of the National Academies, Washington, D.C., 2002, pp. 30–39.
13. Wu, C.-H., C.-C. Wei, D.-C. Su, M.-H. Chang, and J.-M. Ho. Travel Time Prediction with Support Vector Regression. *Proc., 6th IEEE Intelligent Transportation Systems Conference,* Shanghai, China, Oct. 12–15, 2003.
14. Vlahogianni, E. I., J. C. Golias, and M. G. Karlaftis. Short-Term Traffic Forecasting: Overview of Objectives and Methods. *Transport Review,* Vol. 24, 2004, pp. 533–557.
15. Lafferty, J., A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proc., 18th International Conference on Machine Learning,* Williamstown, Mass., June 28–July 1, 2001.
16. Rabiner, L., and B. Juang. An Introduction to Hidden Markov Models. *IEEE ASSP Magazine,* Vol. 3, No. 1, Jan. 1986, pp. 4–16.
17. Kumar, S., and M. Hebert. Discriminative Random Fields: A Discriminative Framework for Contextual Interaction in Classification. *Proc., International Conference on Computer Vision,* 2003, pp. 1150–1159.
18. Liu, Y., J. Carbonell, J. Klein-Seetharaman, and V. Gopalakrishnana. Comparison of Probabilistic Combination Methods for Protein Secondary Structure Prediction. *Bioinformatics,* Vol. 20, No. 17, 2004, pp. 3099–3107.
19. Qin, T., T. Y. Liu, X. D. Zhang, D. S. Wang, and H. Li. Global Ranking Using Continuous Conditional Random Fields. *Proc., Advances in Neural Information Processing Systems,* 2008, pp. 1281–1288.
20. Radosavljevic, V., S. Vucetic, and Z. Obradovic. Continuous Conditional Random Fields for Regression in Remote Sensing. *Proc., European Conference on Artificial Intelligence,* Lisbon, Portugal, Aug. 2010.
21. Coifman, B. Improved Velocity Estimation Using Single Loop Detectors. *Transportation Research Part A,* Vol. 35, No. 10, 2001, pp. 863–880.

The Transportation Network Modeling Committee peer-reviewed this paper.