SHORT-TERM TRAFFIC FLOW RATE FORECASTING BASED ON IDENTIFYING SIMILAR TRAFFIC PATTERNS

Filmon G. Habtemichael, Ph.D.*
Post-Doctoral Research Associate
Civil & Environmental Engineering, Transportation Research Institute at Old Dominion University (ODU)
132 Kufman Hall, Norfolk, VA 23529
fhabtemi@odu.edu
+757-652-1320

Mecit Cetin, Ph.D.
Associate Professor
Director of Transportation Research Institute at ODU
Civil & Environmental Engineering, Transportation Research Institute at Old Dominion University (ODU)
135 Kufman Hall, Norfolk, VA 23529
mcetin@odu.edu
+757-683-6700

* Corresponding author

Abstract

The ability to timely and accurately forecast the evolution of traffic is very important in traffic management and control applications. This paper proposes a non-parametric and data-driven methodology for short-term traffic forecasting based on identifying similar traffic patterns using an enhanced K-nearest neighbor (K-NN) algorithm. Weighted Euclidean distance, which gives more weight to recent measurements, is used as a similarity measure for K-NN. Moreover, winsorization of the neighbors is implemented to dampen the effects of dominant candidates, and rank exponent is used to aggregate the candidate values. Robustness of the proposed method is demonstrated by implementing it on large datasets collected from different regions and by comparing it with advanced time series models, such as SARIMA and adaptive Kalman Filter models proposed by others. It is demonstrated that the proposed method reduces the mean absolute percent error by more than 25%. In addition, the effectiveness of the proposed enhanced K-NN algorithm is evaluated for multiple forecast steps and also its performance is tested under data with missing values. This research provides strong evidence suggesting that the proposed non-parametric and data-driven approach for short-term traffic forecasting provides promising results. Given the simplicity, accuracy, and robustness of the proposed approach, it can be easily incorporated with real-time traffic control for proactive freeway traffic management.

Key words: Short-term traffic forecasting; K-nearest neighbor; K-NN; traffic patterns; Euclidean distance; correlation distance; winsorization; rank exponent; traffic management; non-parametric

1. Introduction

The prime objective of traffic management strategies is to handle road traffic operations up to the highest level of service possible to provide reliable, safer, and greener transportation. In past decades, traffic management has been limited to responsive schemes which react to prevailing traffic conditions. However, with the advancement in technology and the wide deployment of intelligent transportation systems, traffic operators are deploying active traffic management strategies which can dynamically apply alternative strategies proactively in response to predicted traffic conditions. Therefore, the ability to timely, reliably, and accurately forecast the dynamics of traffic over short-term horizons is becoming very important. Short-term traffic forecasting models, therefore, are an integral element of the toolset needed for real-time traffic control and management. Moreover, such tools are important in providing travelers with reliable travel time information, optimizing traffic signals, and deployment of emergency management systems.

Given the importance of predicting the expected volume of traffic ahead of time, considerable amount of research has been focused on the topic (see Van Lint and Van Hinsbergen (2012) and Vlahogianni et al. (2014)). The availability of a vast amount of spatial and temporal traffic data coupled with advancements in statistics and data analysis techniques have created an opportunity to perform short-term traffic forecast with a reasonable prediction accuracy and short processing time. Short-term traffic forecast aims at predicting the evolution of traffic over time horizons ranging from few seconds to few hours (Van Lint and Van Hinsbergen, 2012; Vlahogianni et al., 2014).

In this paper, we present a methodology for short-term traffic forecast based on learning similar traffic patterns as a reference for providing the predictions on future traffic. Similar traffic profiles are identified using an enhanced K-nearest neighbor algorithm based on measurements of a sequence of volume of traffic at 15-minute intervals. For a given prevailing volume profile, K similar profiles (nearest neighbors or candidates) are identified from a large collection of a historic traffic database. The neighboring candidates drawn from the historic database corresponding to the desired forecast time or horizon are aggregated to provide predictions of future flow rate measurements. In this paper, two similarity measures are considered to determine similar traffic patterns, namely: 1) Correlation distance, and 2) Weighted Euclidean distance. To reduce noise while computing the distance measures, the lagging volume profiles are slightly filtered using locally estimated scatter-plot smoothing (loess) technique. Moreover, to dampen the effect of dominant or extreme values of candidate neighbors Winsorization is applied.

As opposed to fitting and optimizing parametric prediction models (e.g., ARIMA), the approach proposed in this paper is fundamentally a data-driven one. The main advantage of this approach is that the predictions are generated based on the observed historical traffic patterns that are discovered from the historical datasets. Moreover, this approach is capable of providing predictions over multiple time steps or a trace forecast over a specified prediction horizon. In this study, predictions are provided over a 15-minute step with a forecast horizon of 6 steps (i.e., predictions were made for one hour and thirty minutes at 15-minute intervals). Considering the simplicity of the proposed approach and a relatively short computation time, it can be easily incorporated with online traffic management strategies to provide short-term traffic forecasts in real-time. The performance of the proposed approach is tested using a wide variety of freeway flow rate datasets collected from different regions and is compared with the works of Guo et al. (2014) which developed several advanced parametric models based on time series analysis reinforced with an adaptive Kalman filter. The proposed methodology is found to outperform the works of Guo et al. (2014) in terms of forecast accuracy; provided that enough samples are available in the archived datasets as explained in the paper. Moreover, comparing the works of this paper with others which applied K-NN approach of forecast, the level of details presented in terms of optimizing the parameters of the enhanced K-NN and testing the robustness of the model under different traffic scenarios corresponding to different datasets collected from different regions and different proportion of missing values is unprecedented.

The remainder of this paper is organized as follows. Following this introductory section, literature review on short-term traffic forecast is presented. This is followed by an in-depth discussion of the proposed methodology for short-term traffic forecast and the corresponding results. Finally, conclusions are drawn and future works are discussed.

2. Literature Review

In the literature, short-time traffic forecast covers prediction of traffic over the time period of a few seconds to few hours in to the future using current and historic measurements of traffic variables (Vlahogianni et al., 2014). According to Van Hinsbergen and Sanders (2007) as well as Van Lint and Van Hinsbergen (2012), the approaches used in short-term traffic forecast can be broadly classified into four categories: Naïve, parametric, non-parametric, and hybrid. Naïve approaches refer to models that provide simple estimate of traffic in the future, e.g., historic averages. Parametric approaches refer to models-based techniques which require a set of fixed parameter values as part of the mathematical or statistical equations they utilize, e.g., analytical models, macroscopic models and models based on time series analysis (e.g., Wang et al., 2006). The majority of these approaches suffer from the assumptions they consider to parameterize the models and were proven to perform relatively poorly under unstable traffic conditions and complex road settings (Vlahogianni et al., 2014). On the other hand, non-parametric approaches are mostly data-driven and apply empirical algorithms to provide the predictions, e.g., approaches based on data analysis and neural network techniques. Such approaches are advantageous as they are free of any assumptions regarding the underlying model formulation and the uncertainty involved in estimating the model parameters. Other short-term traffic models have implemented a hybrid of the above-mentioned approaches (e.g., Szeto et al., 2009). Smith et al. (2002), Lin et al. (2013) and Lippi et al. (2013) have provide a comparative analysis of a few models selected amid many.

The majority of the studies on short-term traffic forecast were conducted using standard statistical techniques such as simple smoothing, complex time series analysis and filtering methods. Application of smoothing for traffic forecast include: kernel smoothing (El Faouzi, 1996), simple exponential smoothing (Ross, 1982), and hybrid exponential smoothing and neural networks (Chan et al., 2012). Others used time series analysis such as Autoregressive Integrated Moving Average (ARIMA) models (Cetin and Comert, 2006; Cools et al., 2009; Hamed et al., 1995; Lee and Fambro, 1999; Moorthy and Ratcliffe, 1988). A variation of the ARIMA model, which is Seasonal ARIMA (SARIMA) models, has also been implemented in many studies (Guo et al., 2008; Lin et al.; Lippi et al., 2013; Szeto et al., 2009; Williams and Hoel, 2003). Szeto et al. (2009) applied combination of cell transmission and SARIMA models. Filtering models, e.g., Kalman filter, have also been applied in short-term traffic forecast (Guo et al., 2014; Okutani and Stephanedes, 1984; Wang and Papageorgiou, 2005; Whittaker et al., 1997). Recently, Chen and Rakha (2014) proposed an algorithm based on particle filters for traffic prediction.

Another line of research on short-term traffic forecast applied neural networks and pattern recognition methods. Some of the works that focused on implementation of neural network and its variations are Innamaa (2005), Li and Chen (2013), Vlahogianni (2007, 2008), Wang and Shi (2013), Zargari et al. (2012), and Zheng et al. (2006). Pattern recognition methods were also applied for short-term traffic forecast, e.g., cluster analysis (Xia et al., 2012), support vector machines (SVM) (Castro-Neto et al., 2009; Wang and Shi, 2013), k-nearest neighbor (Lin et al., 2013; Zhang et al., 2013; Zheng et al., 2006).

For an extended review of the various works on short-term forecasting, various models developed and their technical aspects, the reader is referred to the works of Van Hinsbergen and Sanders (2007), Van Lint and Van Hinsbergen (2012), and Vlahogianni et al. (2004, 2014).

K-NN approach has been previously applied for the purpose of forecasting traffic. A number of researchers have applied K-NN to forecast traffic flow rates (Clark, 2003; Davis and Nihan, 1991; Oswald et al., 2000; Smith and Demetsky, 1997; Smith et al., 2002). Similarly, other researchers applied K-NN for forecasting travel times (Bajaw et al., 2003; Chung and Kuwahara, 2003; Rauscher, 1998; You and Kim, 2000; Robinson and Polak, 2005). The main limitations of the works of these authors are that the simplest form of K-NN was used. In addition, majority of these authors failed to compare their results with parametric models with the exception of Smith et al. (2002).

In summary, considering the fact that there are a wide variety of models developed for short-term traffic forecasts, it can be overwhelmingly difficult to identify the one that is most suitable for specific traffic conditions. Moreover, since the studies were conducted under different settings (such as traffic parameters considered, the nature of datasets and their sources, aggregation intervals, forecast step durations, prediction horizons and performance measures considered), it is very difficult to fairly compare one approach over another. Various researchers have concluded that the performance of non-parametric models is better when compared to parametric models as they are better suited to learn more from the complex data and adapt to its pattern. For example, Van Lint and Van Hinsbergen (2012) suggested that in the context of traffic forecast, non-parametric approach is the first choice as the input and output traffic variables are noisy and the relationship between each other is nonlinear and poorly understood. Pattern recognition-based approaches, a subset of the non-parametric approaches, seem to be more appropriate as they are effective in identifying similar traffic conditions needed to generate a prediction.

3. Methodology

In this research, forecasting is performed by exploiting similarities in traffic patterns. In essence, the future is predicted based on similar profiles from the archived data. This is graphically shown in Figure 1. Suppose it is desired to forecast what the anticipated volume of traffic will be after 2:00 PM on August 06 (shown in solid red line in Figure 1). Having a look at the historic volume profiles, August 06 is found to have similar flow rate profiles with what has been observed on April 04 and May 27 in the past (shown in solid black and blue lines in Figure 1). Hence, the predicted flow rate on August 06 after 2:00 PM can be estimated using some form of aggregation of the flow rates drawn from the similar profiles (shown in broken green line in Figure 1).

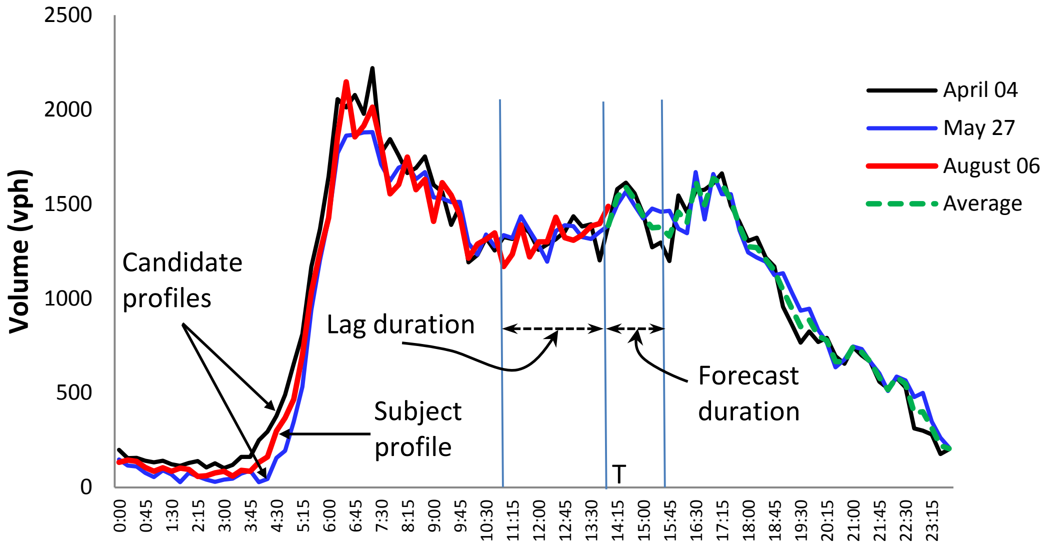

Figure 1 Graphical representation of K-NN-based short-term traffic forecast (data used is from I-169 NB station 737 in MN in the year 2004)

Several terms are defined below before explaining the details of the proposed K-NN method:

Subject profile: The traffic profile corresponding to a specific day for which the forecast is desired to be made (i.e., August 06).

Candidate profiles: Also called nearest neighbors, these are set of traffic profiles with similar pattern to the subject profile selected from a pool of archived dataset (i.e., April 04 and May 27). The number of such candidate profiles considered is referred to as K.

Lag duration: This is the time window considered to determine similarity between the subject and candidate profiles. The number of the sequence of traffic measurements considered in the lag duration is represented by *m*. For example, if the sequence of measurements of flow rate over the past 3 hours is used for matching similar traffic patterns, then the lag duration is 3 hours or *m = 12 (i.e., 3h × 60min/h ÷ 15min aggregation)*.

Lagging part of traffic profile: This is the part of traffic profile used for determining similarity of traffic patterns. Considering that the forecast is being made at time *T*, for a given traffic profile of $x = (x_1, x_2, \dots, x_n)$ and lag duration of *m*, then the lagging part of profile *x* will be the vector $x_T^m = (x_T, x_{T-1}, x_{T-2}, \dots, x_{T-m})$

Size of historic profile: This refers to the magnitude of the archived data, or the search space, from which the candidate profiles are drawn, e.g., archived flow rate profiles of the past one year.

In this paper, K-NN is used as the basic algorithm to identify similar traffic profiles. K-NN is a non-parametric pattern recognition technique which is commonly used for classification and regression purposes. Given an unlabeled object, the algorithm searches for similar objects or neighbors from a search space and assigns a label to the unlabeled object based on the nature of the nearest neighbors. The same concept can also be applied for a sequence of observations, e.g., flow rate measurements. The algorithm identifies the K most similar historic sequences with similar pattern to the one being examined. The combination of the nearest values corresponding to the time step where forecast is desired to be made will be the expected future value of the sequence being examined. The notion of K-NN-based forecast is that the pattern of the sequence of observations is repeated over time. Therefore, if a previous pattern can be identified to be similar to the current pattern, then the subsequent values of the previous sequence can be used to predict the future values of the target sequence (Meade, 2002).

K-NN-based forecast has been applied in other fields as well, e.g., electricity demand forecast (Al-Qahtani and Crone, 2013) and currency exchange forecast (Meade, 2002). A similar principle can be applied in short-term traffic forecast.

In the process of K-NN-based identification of similar traffic conditions and using them as a basis for providing traffic predictions, a few questions need to be answered to specify the required K-NN variables so that the prediction errors are minimal. These include: how to identify similar traffic patterns, how large should the lag duration be, how many nearest neighbors (candidates) should be considered and what mechanism of aggregation should be used to provide an estimate of the future traffic flow rate. Figure 2 shows the methodological procedure adapted in this paper. Detailed discussion of these is provided in the following sections.

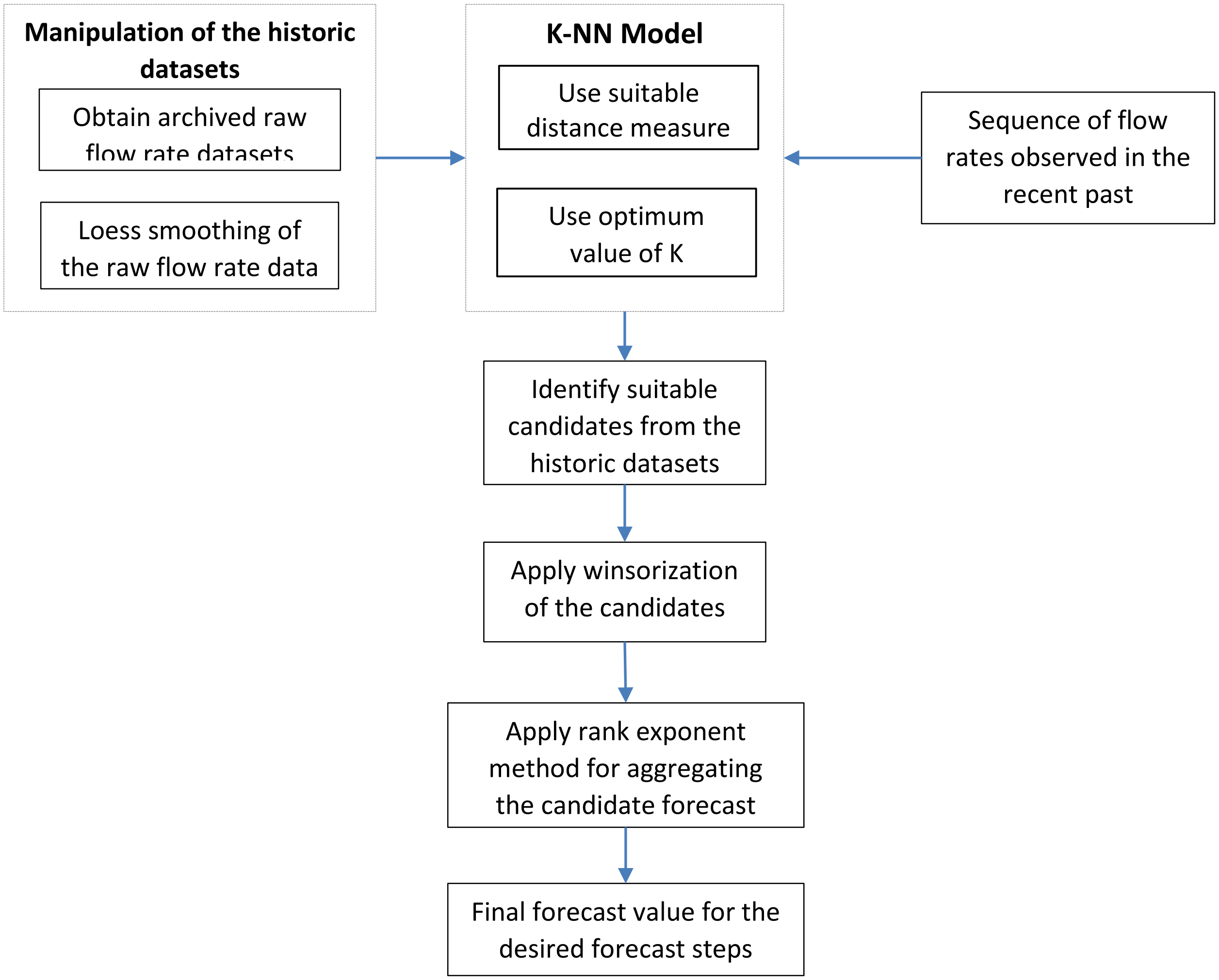
Manipulation of the historic datasets
Obtain archived raw flow rate datasets
Loess smoothing of the raw flow rate data
K-NN Model
Use suitable distance measure
Use optimum value of K
Sequence of flow rates observed in the recent past
Identify suitable candidates from the historic datasets
Apply winsorization of the candidates
Apply rank exponent method for aggregating the candidate forecast
Final forecast value for the desired forecast steps

Figure 2 Flow chart showing the methodological approach

3.1. Measuring similarity between volume profiles

In the area of pattern recognition for time series data, several measures of similarity are commonly used, e.g., correlation distance, cross-correlation distance, Euclidean distance, weighted Euclidean distance, dynamic time warping distance, Fourier transform distance and others (Mori et al., 2014). After analyzing the effectiveness of each measure of similarity for traffic data, two distance measures are taken into consideration: 1) correlation distance and 2) weighted Euclidean distance. The one with the least forecast error is used for prediction purposes.

Correlation distance:

Considering the recent works of Zheng and Su (2014) which claimed that correlation distance is a superior measure of similarity for traffic data, we decided to consider it as a potential measure of similarity. The correlation distance is computed from Pearson's correlation between a pair of numeric time series (Mori et al., 2014) and it measures the degree of dependence of the two time series. Assuming two finite and equidistant series of *x* and *y*, where *x* = (*x₁*, *x₂*,…, *xₙ*) and *y* = (*y₁*, *y₂*,…, *yₙ*), and forecast time *T* and lag duration of *m*, the correlation distance between the lagging profiles of x_T^m and y_T^m is given by Equation 1:

$$D_{(x_T^m, y_T^m)} = 1 - \ Cor(x_T^m, y_T^m) \qquad \text{Equation 1}$$

Where $Cor(x_T^m, y_T^m)$ is the Pearson's correlation between the lagging profiles of x_T^m and y_T^m. $Cor(x_T^m, y_T^m)$ is given by Equation 2:

$$Cor(x_T^m, y_T^m) = \frac{\sum_{i=0}^{m-1}\left(x_{T-i} - \overline{x_T^m}\right)\left(y_{T-i} - \overline{y_T^m}\right)}{\sqrt{\sum_{i=0}^{m-1}(x_{T-i} - \overline{x_T^m})^2 \sum_{i=0}^{m-1}(y_{T-i} - \overline{y_T^m})^2}} \qquad \text{Equation 2}$$

Where: $x_T^m = (x_T, x_{T-1}, x_{T-2}, \dots, x_{T-m})$ and $y_T^m = (y_T, y_{T-1}, y_{T-2}, \dots, y_{T-m})$ are the lagging profiles.
x_{T-i} and y_{T-i} are the i^{th} elements of the lagging profiles x_T^m and y_T^m, respectively.
$\overline{x_T^m}$ and $\overline{y_T^m}$ are the average values of the lagging profiles x_T^m and y_T^m, respectively.

Weighted Euclidean Distance

Euclidean distance is the distance of a straight-line connecting two points and is the most common distance measure. For a sequence of points, x_T^m and y_T^m, it is obtained by aggregating the distance between the corresponding data points in the sequence as shown in Equation 3.

$$D_{(x_T^m, y_T^m)} = \sqrt{\sum_{i=0}^{m-1}(x_{T-i} - y_{T-i})^2} \qquad \text{Equation 3}$$

In this paper, weights are introduced to Euclidean distance measures according to the importance of the measurements in the sequence. This is done to give more weight to recent flow rate measurements and less to older ones so that the traffic pattern closest to the prevailing conditions are accurately identified. Weighted Euclidean distance is, therefore, given as shown in Equation 4.

$$D_{(x_T^m, y_T^m)} = \sqrt{\sum_{i=0}^{m-1} w_i \ \times \ (x_{T-i} - y_{T-i})^2} \qquad \text{Equation 4}$$

Where w_i ($0 < w_i < 1$) represents the weights assigned to the values in the lagging profiles sequence of x_T^m and y_T^m. In this study, the weights were distributed linearly according to the recentness of the values.

3.2. *Reducing noise in the lagging profiles*

The flow rate data were aggregated every 15 minutes and some noise was observed in the data. This has a negative impact on measuring the similarity of given traffic profiles. To reduce the noise, the parts of the flow rate profiles that are used to compute similarity measures are slightly smoothed using locally estimated scatterplot smoothing (loess) with a span of 0.2. Note that only the lagging parts of the traffic profiles in the search space are smoothed and not the parts that are used to provide the forecasts. Loess is advantageous as it is a non-parametric regression technique which is highly flexible and attempts to capture the pattern of the data without any assumption on the nature of the raw data (Cleveland and Devlin, 1988). Figure 3 shows raw and loess smoothed flow rate data for a typical day.

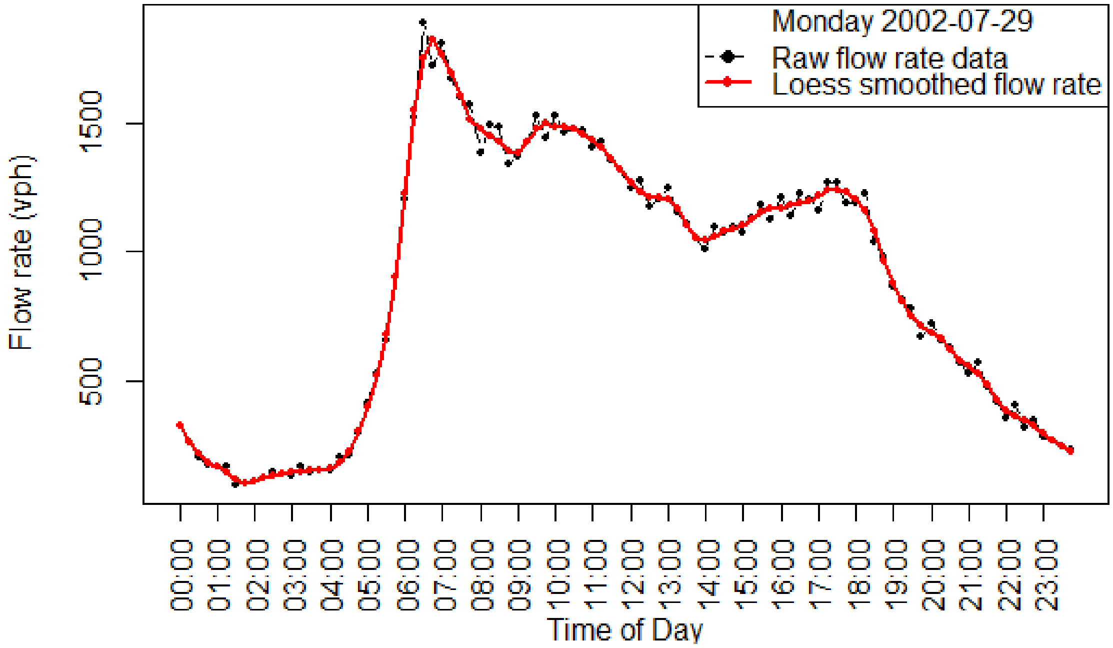

Figure 3 A sample of raw and loess smoothed flow rate data

3.3. Damping the effect of extreme candidate values

Predictions provided from k-NN-based approach depend on the values of the candidates (or nearest neighbors). Extreme value of one candidate can be dominant over the values of the rest of the candidates and thus provides biased predictions. To dampen the effect of such extreme values, winsorization of the candidates is applied. Winsorization is the process of replacing the smallest and largest values of the candidate profiles with the values closest to them. Winsorization is particularly useful when working with traffic data that are affected by exogenous factors such as incidents and adverse weather conditions. Since the incident and weather records corresponding to the traffic datasets that are used in this paper were not provided, winsorization plays a vital role in damping the effect of extreme candidate values.

Mathematically, winsorization is shown in Equation 5. Assuming the candidate values are given by c, where $c = (c_1, c_2, \ldots, c_K)$, and its elements are arranged in an increasing order, then the winsorized candidate values (c_i^w) will be:

$$c_i^w = \begin{cases} c_{(i+1)} & if \qquad c_i = min(c) \\ c_{(i)} & if \qquad \min(c) < c_i < max(c) \\ c_{(K-1)} & if \qquad if\ c_i = \max(c) \end{cases}$$ Equation 5

Where: c_i is the value of the *i*th candidate
min(c) is the minimum value of the candidates
max(c) is the maximum value of the candidates
K is the number of candidates
c_i^w is the winsorized value of the *i*th candidate

3.4. Weight assignment to the candidate values

The notion of assigning weight is to systematically combine the forecasts implied by the candidate values, i.e., the influence of the candidates on the forecasted value are adjusted according to their proximity to the subject profile. There are a number of weight assignment methods which include: 1) Equal weights (all candidates are given the same weight, i.e., simple arithmetic average), 2) Inverse distance weights or Shepard's method (assumes the influence of a candidate profile drops-off with increase in distance from the subject profile), and 3) Rank-based weights (weights are assigned depending on rank of candidates when sorted according to increasing order of distances from the subject profile). Recently, the authors have shown that the performance of rank-based weight assignment method is better when compared to the equal or inverse distance weighting methods in identifying similar traffic patterns (Habtemichael et al., 2015). Therefore, Rank-Exponent method of weight assignment is used in this paper. Rank-Exponent method is advantageous as it provides some degree of flexibility in the way weights are assigned by adjusting the weight dispersion measure as shown in Equation 6. The value of *z* is set to be 2 as indicated by previous work of the authors (Habtemichael et al., 2015).

$$\boldsymbol{W_i} = \frac{(\boldsymbol{K} - \boldsymbol{r_i} + \mathbf{1})^{\boldsymbol{z}}}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{K}} (\boldsymbol{K} - \boldsymbol{r_i} + \mathbf{1})^{\boldsymbol{z}}} \qquad \text{Equation 6}$$

Where: r_i is the rank of the *i*th candidate
K is the total number of candidates
z is weight dispersion measure

3.5. Measuring performance of the proposed short-term traffic forecast method

Three measures are used as performance indicators for accuracy of the proposed method of short-term forecast and they are: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) as shown in Equation 7 through Equation 9. Such multiple measures of performance provide an in-depth understanding on the nature of the forecast errors. For example, MAPE provides the forecast error in terms of percentage difference between the observed and predicted flow rates, MAE and RMSE provide the forecast error in terms of differences in the count of vehicles. Higher values of RMSE than MAE indicate there is some variation in the magnitude of the forecast errors in the prediction made.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |F_i - O_i| \qquad \text{Equation 7}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{F_i - O_i}{O_i}\right| \times 100\%$$ Equation 8

$$RMSE = \frac{1}{n}\sqrt{n\sum_{i=1}^{n}(F_i - O_i)^2}$$ Equation 9

Where: F_i is the i^{th} forecast value
O_i is the i^{th} observed value
n is the number of samples

The performance of the proposed short-term traffic forecast is investigated in two ways: 1) by traffic flow level at the time of forecast, and 2) by time of day when the forecast is made. This allows examining the accuracy of the forecast from multiple perspectives. Forecast accuracy by level of traffic is given by creating volume bins with an increment of 500 veh/h/ln. These volume bins can be approximately translated into Levels Of Service (LOS) as shown in in Table 1, assuming uncongested flow conditions. Such category of volume bins has been used by Guo et al. (2014). Similarly, forecast accuracy by time of day refers to the hour of the day when the forecast is made and continues from 12:00 AM to 11:00 PM with one-hour intervals.

Table 1 Definition of traffic levels or traffic volume groups (adopted from Guo et al., 2014)

Volume groups	Group description	Approximate level of service (LOS)
Group 1	≥ 0 and < 500 veh/h/ln	LOS A
Group 2	≥ 500 and < 1000 veh/h/ln	LOS B
Group 3	≥ 1000 and < 1500 veh/h/ln	LOS C
Group 4	≥ 1500 and < 2000 veh/h/ln	LOS D
Group 5	≥ 2000 veh/h/ln	LOS E

4. Experiments

4.1. Data used

Multiple datasets collected from different regions, same as those used in Guo et al. (2014), are used in this paper (these are graciously provided by Drs. Williams and Guo). This includes 12 datasets from the motorways of United Kingdom (UK) and 24 datasets from freeways within the United States (Maryland (MD), Minnesota (MN) and Washington (WA) each with 6, 12 and 6 datasets, respectively). This allows testing the proposed methodology under a variety of driving environments and traffic conditions. Previously, it was shown that 15-minute data aggregation is ideal for short-term traffic forecast (Guo et al., 2014). Therefore, the datasets are aggregated at 15-minute intervals and missing data are imputed using screening procedures and SARIMA(1,0,1)(0,1,1)$_{672}$ model (672 refers to the number of one-week flow rate measurements at 15-minute intervals, i.e., 1x7x24x4=672). Brief description of the datasets is shown in

Table 2 (the reader is referred to Guo et al. (2012) and Guo et al. (2014) for more discussion on the data used).

Table 2 Brief description of the datasets used (adopted from Guo et al., 2014))

Region	Highway	Station	No. of lanes	Start	End	No. of months
UK	M25	4762a	4	9/1/1996	11/30/1996	3 months
UK	M25	4762b	4	9/1/1996	11/30/1996	3 months
UK	M25	4822a	4	9/1/1996	11/30/1996	3 months
UK	M25	4826a	4	9/1/1996	11/30/1996	3 months
UK	M25	4868a	4	9/1/1996	11/30/1996	3 months
UK	M25	4868b	4	9/1/1996	11/30/1996	3 months
UK	M25	4565a	4	1/1/2002	12/31/2002	12 months
UK	M25	4680b	4	1/1/2002	12/31/2002	12 months
UK	M1	2737a	3	2/13/2002	12/31/2002	11 months
UK	M1	2808b	3	2/13/2002	12/31/2002	11 months
UK	M1	4897a	3	2/13/2002	12/31/2002	11 months
UK	M6	6951a	3	1/1/2002	12/31/2002	12 months
MD	I270	2a	3	1/1/2004	5/5/2004	4 months
MD	I95	4b	4	6/1/2004	11/5/2004	6 months
MD	I795	7a	2	1/1/2004	5/5/2004	4 months
MD	I795	7b	2	1/1/2004	5/5/2004	4 months
MD	I695	9a	4	1/1/2004	5/5/2004	4 months
MD	I695	9b	4	1/1/2004	5/5/2004	4 months
MN	I35W-NB	60	4	1/1/2000	12/31/2000	12 months
MN	I35W-SB	578	3	1/1/2000	12/31/2000	12 months
MN	I35E-NB	882	3	1/1/2000	12/31/2000	12 months
MN	I35E-SB	890	3	1/1/2000	12/31/2000	12 months
MN	I69-NB	442	2	1/1/2000	12/31/2000	12 months
MN	I69-SB	737	2	1/1/2000	12/31/2000	12 months
MN	I35W-NB	60	4	1/1/2004	12/31/2004	12 months
MN	I35W-SB	578	3	1/1/2004	12/31/2004	12 months
MN	I35E-NB	882	3	1/1/2004	12/31/2004	12 months
MN	I35E-SB	890	3	1/1/2004	12/31/2004	12 months
MN	I69-NB	442	2	1/1/2004	12/31/2004	12 months
MN	I69-SB	737	2	1/1/2004	12/31/2004	12 months
WA	I5	ES-179D_MN_Stn	4	1/12004	6/29/2004	6 months
WA	I5	ES-179D_MS_ Stn	3	1/12004	6/29/2004	6 months
WA	I5	ES-130D_MN_ Stn	4	4/1/2004	9/30/2004	6 months
WA	I5	ES-179D_MS_ Stn	4	4/1/2004	9/30/2004	6 months
WA	I405	ES-738D_MN_ Stn	3	7/1/2004	12/29/2004	6 months
WA	I405	ES-738D_MS_ Stn	3	7/1/2004	12/29/2004	6 months

As indicated previously, the accuracy of the proposed K-NN model is compared to the advanced time-series models recently developed by Guo et al. (2014). Their study documents aggregated results of various adaptive methods (see Section 4.4) for all days in the datasets except the first week in each dataset which is used for model training. To be able to compare the K-NN results directly with those of Guo et al. (2014) and to be consistent with their numerical experiments, the K-NN method is applied to all days in the given datasets with the exception of those in the first week. However, when searching for similar days for any subject day, the entire dataset (except the subject day), including both past and future days, is used. For example: assuming a forecast is being made for October 1, 1996, for station 4762a on UK motorway M25 (see Table 2), then the search space considered is from September 1 to November 1, 1996. Obviously, some days in the search space are future days relative to October 1, 1996, which is not going to be realistic when performing forecasting in real-world applications. However, this issue is negligible for the analyses presented here since it is assumed that large archived data will be available when using K-NN for forecasting in the real-world. Furthermore, to justify the use of both future and past days in K-NN method, for selected days in yearly datasets, forecasts are performed based only on past observations and mixed observations (past and future days). It is found that there is no statistically significant difference between the predictions errors obtained under these two scenarios.

4.2. Variable estimation for K-NN-based traffic forecast

For the proposed K-NN algorithm several variables have to be determined beforehand so that the forecast error is as small as possible. These variables include the selection of a suitable distance measure, lag duration, and number of nearest candidates. In data analysis, it is customary to train and test a model on separate datasets to avoid overfitting. In this paper, a three-fold cross validation technique of model training is used, i.e., only one third of the datasets (12 datasets representing all the regions) are used as training datasets for estimation of variables for the K-NN-based short-term traffic forecast model, i.e., optimum distance measure to be used, and identifying suitable lag duration and number of candidates (K). Once these optimum parameters are identifies using the 12 datasets, the enhanced K-NN model is applied to all 36 dataset and the overall forecast accuracy is examined.

Identifying the optimum distance measure

As discussed previously, two distance measures were considered for identifying similar traffic patterns. Figure 4 shows the average forecast errors corresponding to the aforementioned distance measures, namely: 1) Correlation distance, and 2) Weighted Euclidean distance. Ten nearest neighboring candidates with lag duration of one hour are used for this purpose (this is discussed later in detail). It can be easily observed that weighted Euclidean distance outperforms correlation distance as the forecast errors are significantly lower. Therefore, weighted Euclidian distance is used throughout the remainder of this paper.

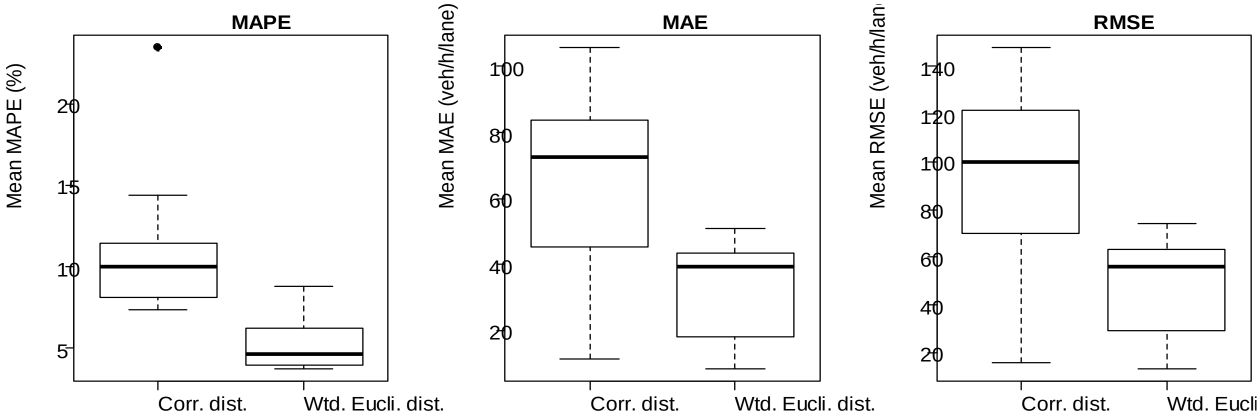

Figure 4 Comparison of Correlation and Weighted Euclidean distances

Identifying suitable lag duration and number of candidates

It is very important that optimal lag duration and number of candidates are used to minimize the forecast error. Lag duration affects the performance of the K-NN based traffic forecast as it is the main variable that identifies similar traffic patterns. A series of lag durations are considered in this study, ranging from just one hour up to 23 hours. Intuitively, shorter lag durations are suitable for short-term traffic forecast while relatively longer durations are suitable for long-term traffic forecast. Another variable that affects the accuracy of the proposed forecast method is selection of the number of candidates. In this paper, a wide variety of numbers of candidates are considered, ranging from just one candidate to 23. The candidates are later aggregated to predict the flow in the next time step using Rank-Exponent method of aggregation (see Equation 6).

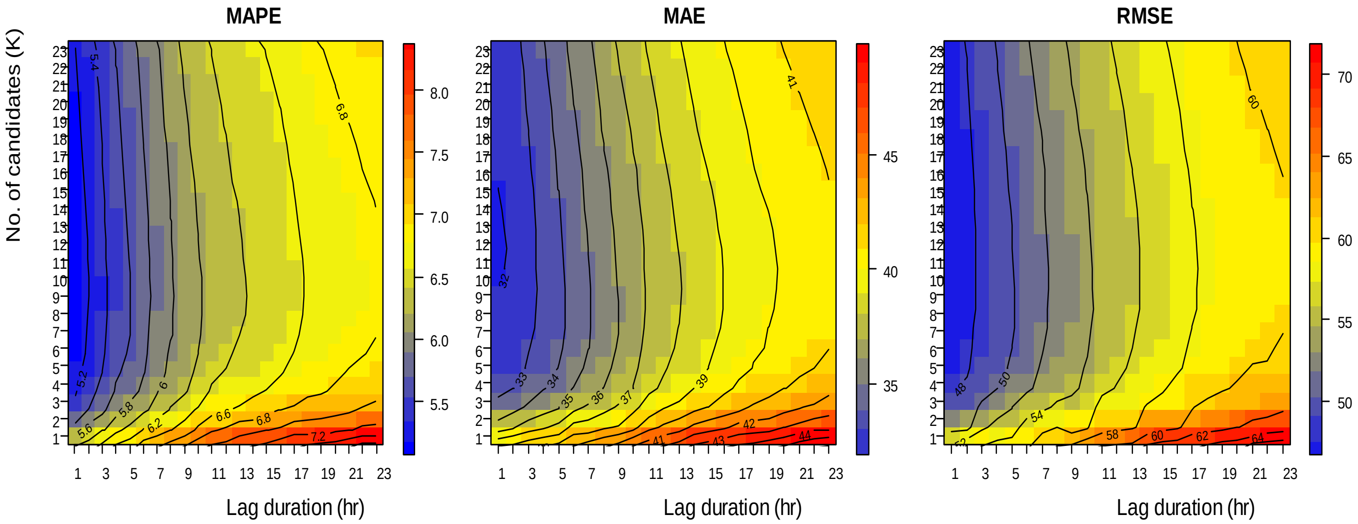

Figure 5 Impact of lag duration and number of candidates on forecast error

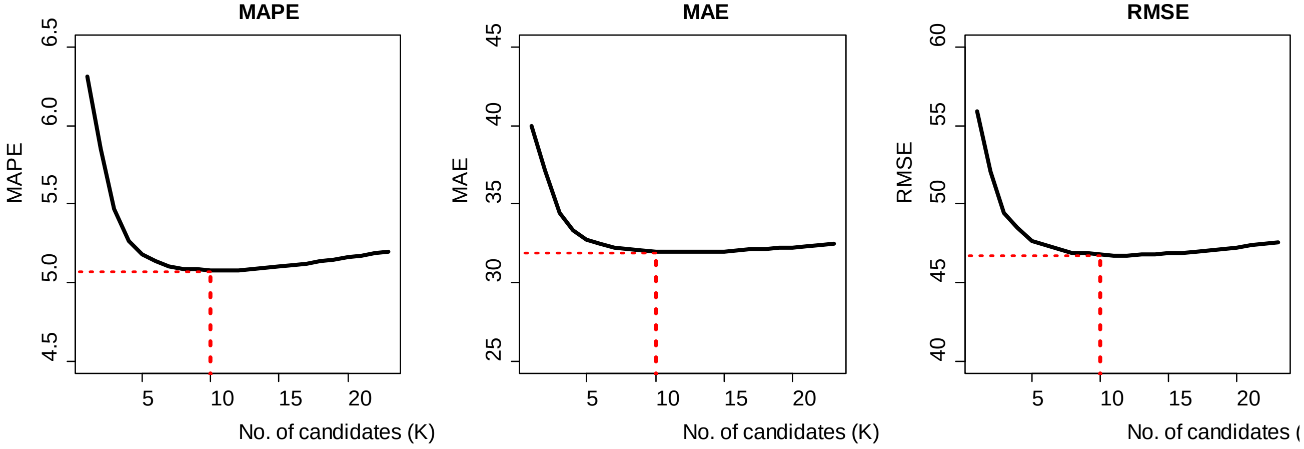

Figure 6 Optimum number of candidates given the lag duration is one hour

The impact of lag duration and number of candidates considered on forecast accuracy in terms of the three error criteria are shown in Figure 5. It can be observed that with increase in lag duration, forecast errors increase. This shows that the optimal lag duration for identifying similar traffic patterns should be of relatively short duration; in our case one hour lag duration is found to be most suitable. Similarly, the number of candidates selected affects the forecast accuracy, although to a lesser degree when compared to the effect of lag duration. A single candidate provides a poor estimate of the future traffic flow rates. With increase in the number of candidates considered, the forecast errors decrease and then start to increase slightly. Figure 6 shows the impact of number of candidates used on the forecast accuracy when lag duration is kept at one hour. The optimum number of candidates to be considered in the proposed K-NN-based traffic forecast is found to be ten.

4.3. Accuracy of forecast by level of traffic and time of day

Following the study by Guo et al. (2014), the accuracy of the proposed forecasting method is examined by level of traffic and by time of day. Figure 7 show the performance of the traffic forecast using the proposed approach in terms of MAE, MAPE and RMSE by level of traffic and time of the day. The box plots show the spread of the forecast errors and the red solid line represents the mean of the errors.

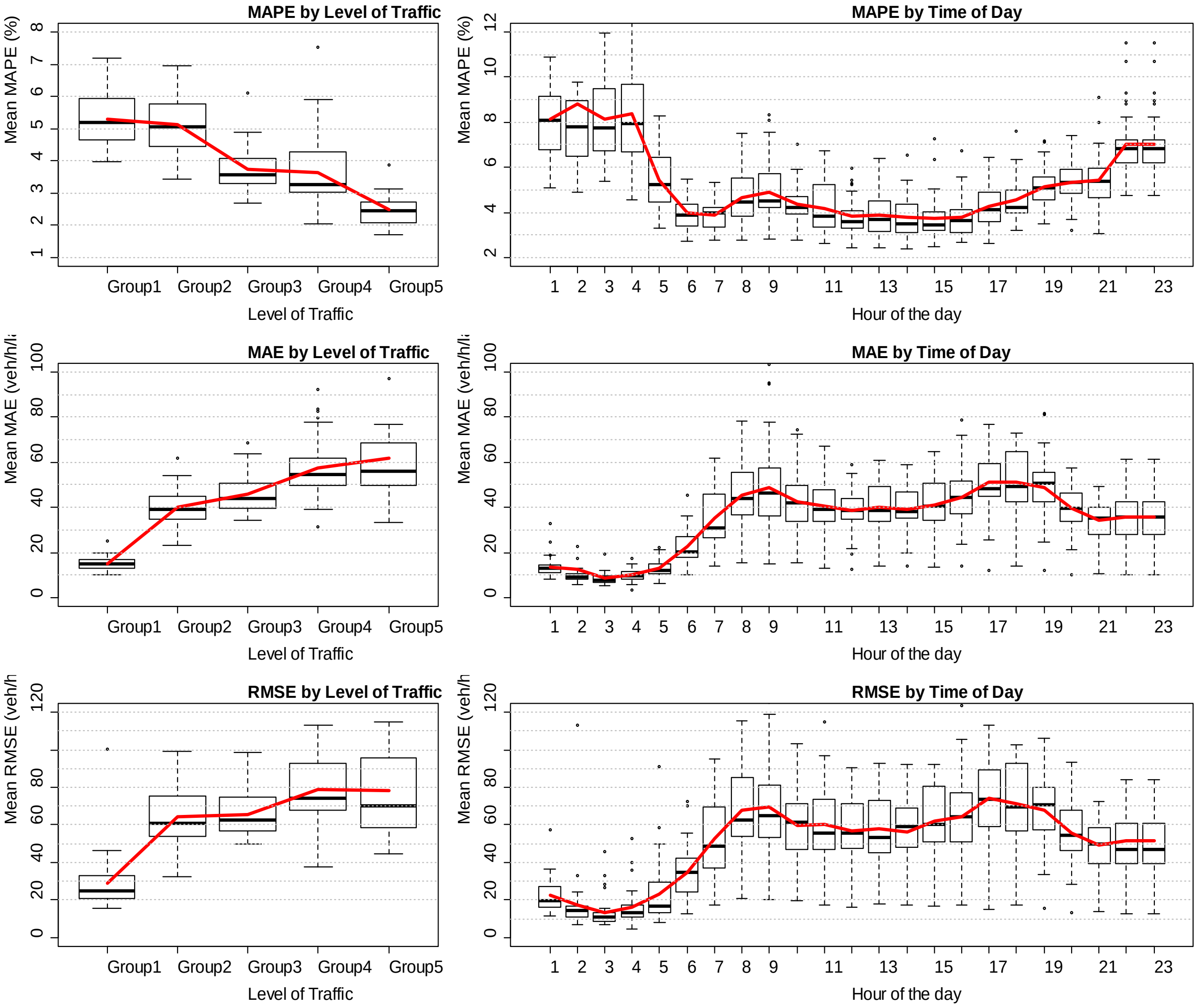

Figure 7 Forecast errors for enhanced K-NN by level of traffic and time of the day

When the nature of errors corresponding to level of traffic is examined, a consistent improvement in forecast accuracy is observed with an increase in the level of traffic during the forecast time. This is evident by consistent reduction in MAPE. MAPE provides a better sense of forecast accuracy as the errors are examined in terms of percentage deviations from the observed value. In other words, the forecast errors are normalized according to the magnitude of the observed values. As expected, with increase in level of traffic, MAE and RMSE increased slightly. This is because of the fact that both MAE and RMSE consider only the magnitude of deviations of the forecasted values from the observed ones.

Similarly, when the nature of the errors corresponding to the time of day is examined, the forecast accuracy, in terms of MAPE, during off-peak hours is relatively lower when compared to peak-hours. The values of MAE and RMSE are low during late night and early morning hours because the observed traffic volumes are low.

Generally speaking, having a look at the spread of the whiskers in Figure 7, it can be said that the proposed enhanced K-NN model provides reliable and accurate forecasts of traffic flow rates. Forecast of traffic flow rates are more important and needed during high levels of traffic (i.e., peak-hours). For example, forecast of traffic in peak hours can be used to predict flow breakdown which is usually followed by a drop in capacity resulting in the worst freeway operations during the time capacity is needed the most. Therefore, a reliable and relatively accurate forecast of traffic during congested traffic conditions can be used as a decision support tool for traffic operators for implementing an alternative traffic management strategy to avoid flow breakdowns. Considering the forecast accuracy of the proposed methodology is high during peak hours, it can be said that its contribution is expected to be high in managing the traffic proactively.

4.4. Comparing the results with those of Guo et al. (2014)

To evaluate the performance of the proposed method, the enhanced K-NN results are compared with those from the recently published works of Guo et al. (2014). For the comparison to be fair, common datasets and measure of performance of traffic forecasts are used. Guo et al. (2014) worked on short-term traffic forecast using a variety of models based on time series analysis. The models they employed include:

1) **EXPRW** employs a seasonal exponential smoothing and uses random walk to capture the pattern as well as the local variations of volume profiles,
2) **BATCH** uses ARIMA to process the SARIMA(1,0,1)(0,1,1)$_{672}$ model and AUTOREG to process the GARCH(1,1) model after transforming the data series by square root transformation,
3) **KF** uses seasonal exponential smoothing and two standard Kalman filters for processing the SARIMA(1,0,1)(0,1,1)$_{672}$ + GARCH(1,1) structure, and
4) **AKF** uses seasonal exponential smoothing and two adaptive Kalman filters for processing the SARIMA(1,0,1)(0,1,1)$_{672}$ + GARCH(1,1) structure.

All four of the parametric models listed above are included in Guo et al. (2014). For more discussion on EXPRW, BATCH, KF and AKF, the reader is referred to Guo et al. (2012) and Guo et al. (2014).

Overall, the proposed approach is able to provide a forecast of traffic flow rates with lower prediction errors. Figure 8 and Figure 9 show the forecast errors from the proposed approach (K-NN), EXPRW, BATCH, KF, and AKF both by level of traffic and time of the day.

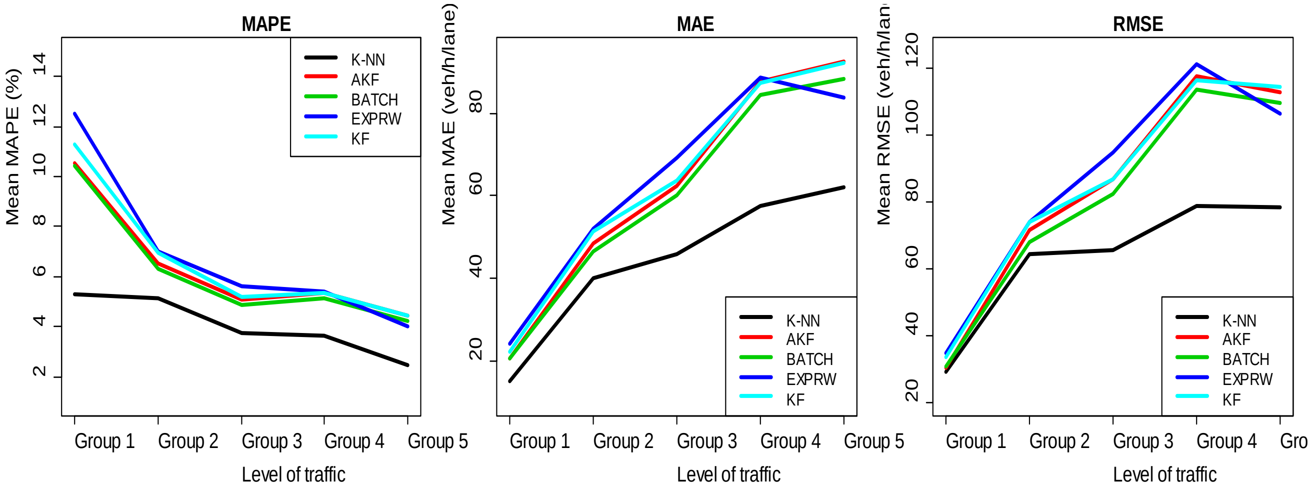

Figure 8 Comparison of forecast with works of Guo et al. (2014) by level of traffic

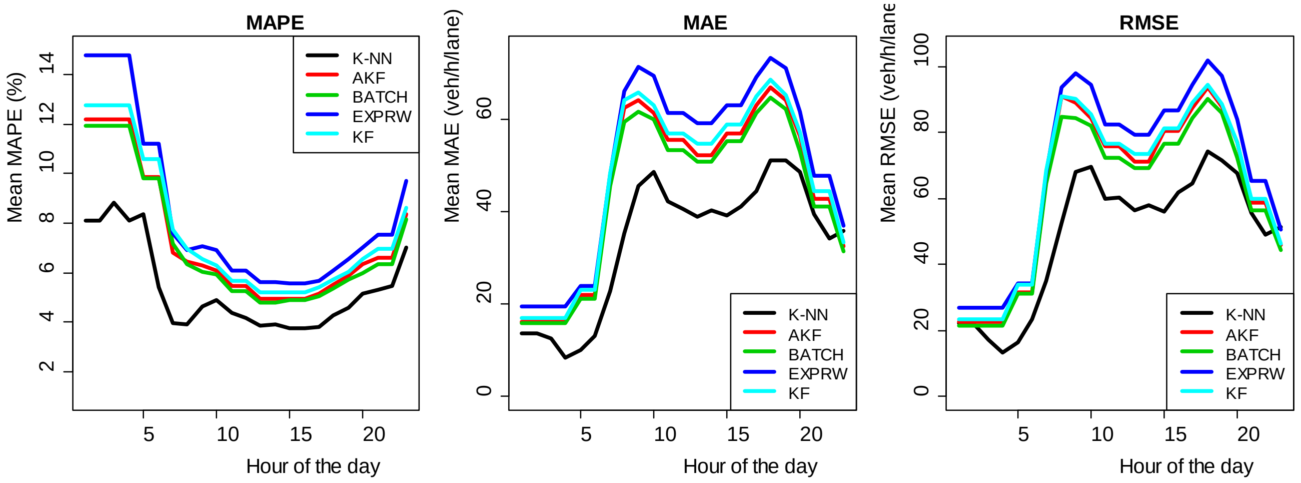

Figure 9 Comparison of forecast errors with works of Guo et al. (2014) by time of day

While more detailed comparisons are given below, the average percentage reductions in forecast errors relative to the works of Guo et al. (2014) are found to be 29% for MAPE, 28% for MAE and 23% for RMSE, which are quite significant. Figure 10 shows the mean forecast errors and their variations for each method of short-term traffic prediction. Significance of the changes in the mean forecast errors between the proposed enhanced K-NN method and AKF, BATCH, EXPRW, and KF is examined using a one-tailed paired t-test. The null hypothesis for the paired t-test is that there is no difference in the forecast error between the proposed methods and those developed by Guo et al. (2014), while the alternative hypothesis is that the reduction in the forecast error is greater than zero. The results indicate that the reductions in forecast errors are found to be very statistically significant and thus the null hypothesis is rejected. Table 3 shows a summary of the comparison of the forecast errors and their statistical significance.

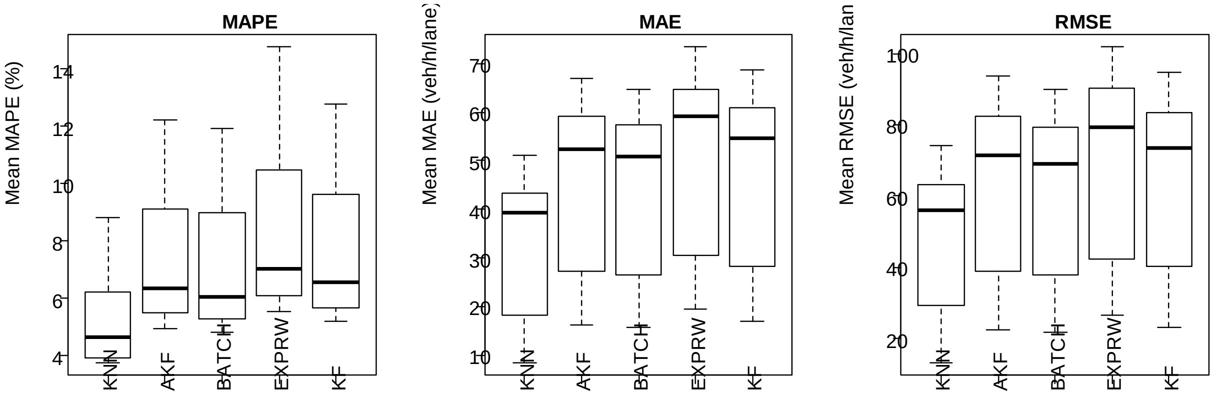

Figure 10 Comparison of mean forecast errors for each method of short-term traffic forecast

Table 3 Summary of comparison of forecast methods (95% confidence level, alpha = 0.05 and t-critical = 1.717)

Forecast Error	Comparison	Percentage reduction	t Stat	P-value	Difference statistically significant
MAPE	KNN vs AKF	27%	-8.019	2.84E-08	Yes
	KNN vs BATCH	25%	-7.335	1.21E-07	Yes
	KNN vs EXPRW	36%	-8.093	2.44E-08	Yes
	KNN vs KF	31%	-8.384	1.34E-08	Yes
MAE	KNN vs AKF	26%	-7.807	4.41E-08	Yes
	KNN vs BATCH	23%	-7.216	1.57E-07	Yes
	KNN vs EXPRW	33%	-10.30	3.49E-10	Yes
	KNN vs KF	29%	-8.366	1.39E-08	Yes
RMSE	KNN vs AKF	22%	-6.700	4.91E-07	Yes
	KNN vs BATCH	19%	-5.971	2.61E-06	Yes
	KNN vs EXPRW	29%	-9.224	2.57E-09	Yes
	KNN vs KF	24%	-7.122	1.92E-07	Yes

The reason that the proposed K-NN method of forecasting traffic flow rate provided better results than the other parametric models is that the model was creatively enhanced to optimize it performance. Moreover, the enhanced K-NN model takes advantage of the big historic datasets which enables it to flexibly adjust itself to adapt to any given traffic conditions as forecasts are drawn from previously observed traffic conditions. It has to be noted that the parametric models (i.e., EXPRW, BATHC, KF and AKF) that were compared with the enhanced K-NN model are not simplistic either.

Comparing the performance of the proposed forecasting method with the works of other researchers is challenging as the experiments are designed in different environments. Nevertheless, recent studies which focused on short-term traffic forecast by Zheng and Su (2014) and Wang et al. (2014) report mean MAPE of about 13.2% and 8.2%, respectively, while the mean MAPE for the proposed method is 5.3%. Zheng and Su (2014) is based on a K-NN method supported by constrained linearly sewing principle component algorithm while Wang et al. (2014) implement an advanced Bayesian combination method. Other studies which applied simple K-NN models for forecasting flow rates reported average MAPE of 8% (Smith and Demetsky, 1997), 9.5% (Smith et al., 2002), 10% (Clark, 2003) and 11% (Oswald et al., 2000). These results are significantly higher than what is reported in this paper.

4.5. Performance of the proposed approach for multiple forecast steps

The proposed algorithm can provide forecast over single step (*h=1*) or a trace of forecasts covering multiple steps, *h = (1, 2,…, H)*. In this study, 6 step predictions covering a forecast horizon of one hour and 30 minutes are provided as shown in Figure 11 and Figure 12. As expected, with an increase in prediction steps the forecast errors are observed to slightly increase. For each prediction step, the average increases in forecast errors of MAPE, MAE and RMSE are found to be 7%, 9% and 9%, respectively. Lag duration of one hour and 10 candidate values are used for the multiple step predictions. Note that Figure 11 and Figure 12 show the accuracy of the trace forecasts for one hour and 30 minutes into the future and all are given at the current time. The accuracy of the multiple forecast steps can be further improved by updating the trace forecasts at the end of the next time step. Since the work of Guo et al. (2014) focused only on one step prediction, we were unable to compare the performance of multiple time step forecast obtained from the enhanced K-NN with the parametric models of EXPRW, BATHC, KF and AKF.

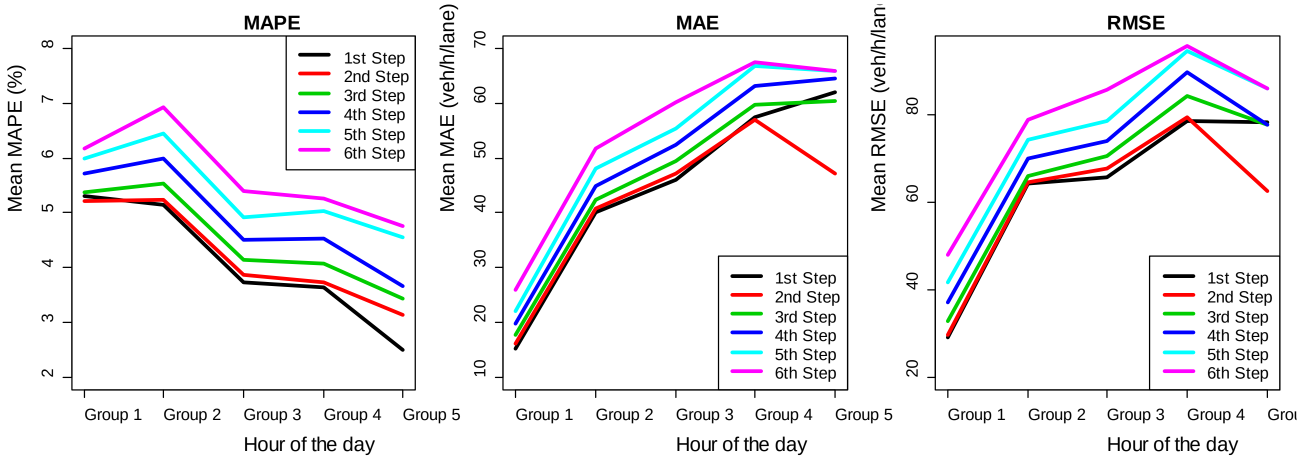

Figure 11 Forecast errors for multiple forecast steps by level of traffic

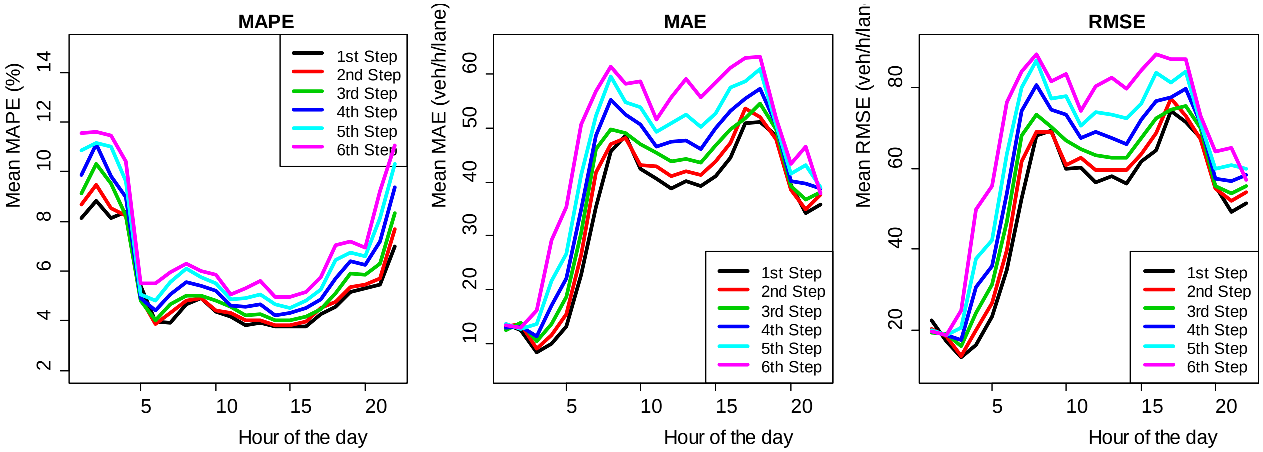

Figure 12 Forecast errors for multiple forecast steps by hour of the day

4.6. Performance of K-NN-based short term traffic forecast under missing values

Missing values due to detector faults or data transmission and storage issues are a common problem in traffic datasets. Therefore, robustness of the forecast models to missing data is very important, especially for real-time traffic application, otherwise the models may provide biased predictions. Some models do not even function under missing values, let alone providing biased predictions. Even though missing values can be imputed, imputing traffic measurements is not straightforward as traffic variables constitute temporal and spatial variations and the relationship among them is not linear. The advantage of the proposed approach of short-term traffic forecast is that it doesn't require the missing values to be imputed.

Some entries of the datasets are deliberately deleted to test the robustness of the proposed forecasting approach under missing values. Different proportions of missing data are considered and the forecast errors are compared with that of complete datasets. The proportion of missing data considered were: 5%, 10%, and 15%.

Measuring similarity under missing values

Whenever the distance between two flow rate profiles containing missing values is being computed, the sequence of the entries with missing values are not considered for distance computation. In other words, only sequence of entries with valid values is considered for distance computation. To compensate for the reduced number of entries in the volume profiles, the weighted Euclidean distance metric was normalized ($D'_{(x_T^m, y_T^m)}$) according to the original and the remaining number of entries as shown in Equation 10.

$$\boldsymbol{D'_{(x_T^m, y_T^m)} = \sqrt{\sum_{i=0}^{m-1} w_i \times (x_{T-i} - y_{T-i})^2} \times \sqrt{\frac{N_O}{N_V}}} \qquad \text{Equation 10}$$

Where: $D'_{(x_T^m, y_T^m)}$ is the normalized distance

N_O is the number of original entries in the lagging parts of the volume profiles

N_V is the number of valid entries in the lagging parts of the volume profiles

The results of the forecast errors under various missing proportions by level of traffic and time of the day are shown in Figure 13 and Figure 14, respectively. Generally speaking, with an increase in missing values, the forecast errors are observed to increase accordingly. However, the magnitude of the increase in the forecast error is fairly small. On average, the forecast error in terms of MAPE, MAE and RMSE under the complete data set was found to be 4.9%, 31 veh/h/ln and 44 veh/h/ln respectively. Under missing values of proportions 5%, 10% and 15%, MAPE slightly increased to 5.3%, 5.5% and 6.9%, respectively. Similarly, MAE increased to 33 veh/h/ln, 35 veh/h/ln and 45 veh/h/ln while RMSE slightly increased to 48 veh/h/ln, 49 veh/h/ln and 50 veh/h/ln under missing values of proportions 5%, 10% and 15%, respectively. The fact that the forecast errors didn't increase significantly with increase in proportion of missing values indicates that the enhanced K-NN model provides unbiased results under incomplete datasets.

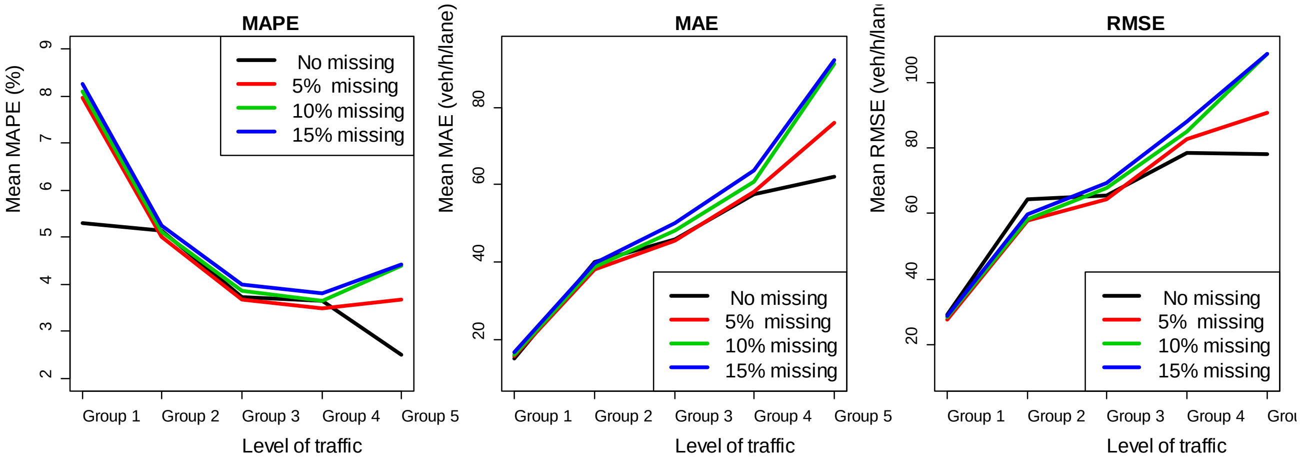

Figure 13 Impact of missing data on forecast errors by level of traffic

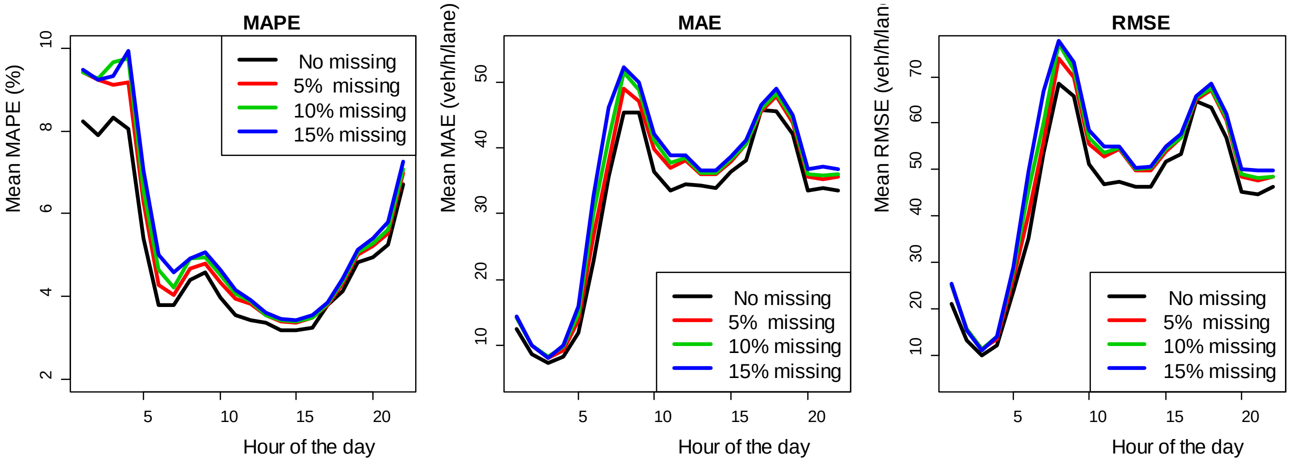

Figure 14 Impact of missing data on forecast errors by hour of the day

4.7. Impact of the size of historic datasets used on forecast accuracy

Despite the simplicity and effectiveness of K-NN in many applications, its performance heavily depends on the size of the training data or the search space from which the neighbors are identified. Likewise, in K-NN-based traffic forecast, the bigger the size of the historic traffic data, the higher the chances of getting candidates with very similar patterns to the subject profile. Therefore, the performance of the proposed enhanced K-NN method of forecast under limited data size is tested and compared with the models AKF, BATCH, EXPRW, and KF. To demonstrate this, two scenarios are developed where the search space from which candidate values are drawn is set to be only: 1) Two Weeks, and 2) One Month as shown in Figure 15. Note that all the enhancements are applied to the Two Week and One Month K-NN models except that the search space is limited to two weeks and one month, respectively.

The results show that small search spaces, e.g., two weeks, provides relatively larger forecast errors. Significance of the difference in forecast error is tested using paired t-test and is found to be statistically significant in the majority of the cases. Moreover, the variability of the forecast errors is large as indicated by the large gap between the upper and lower fences corresponding to the scenario where only two weeks data is used. This implies that, for traffic forecasting purposes, it is better to use models that are based on time series analysis, e.g., such as those developed by Guo et al. (2014), when the available data size is fairly small.

Besides the forecast accuracy, the size of the search space is inversely related to computation speed, i.e., forecasts drawn from large search space require longer computation time. Moreover, larger search spaces require traffic management centers to archive historic data for longer durations. However, considering the advances in modern computing power and data archiving technologies, this is not expected to be a concern.

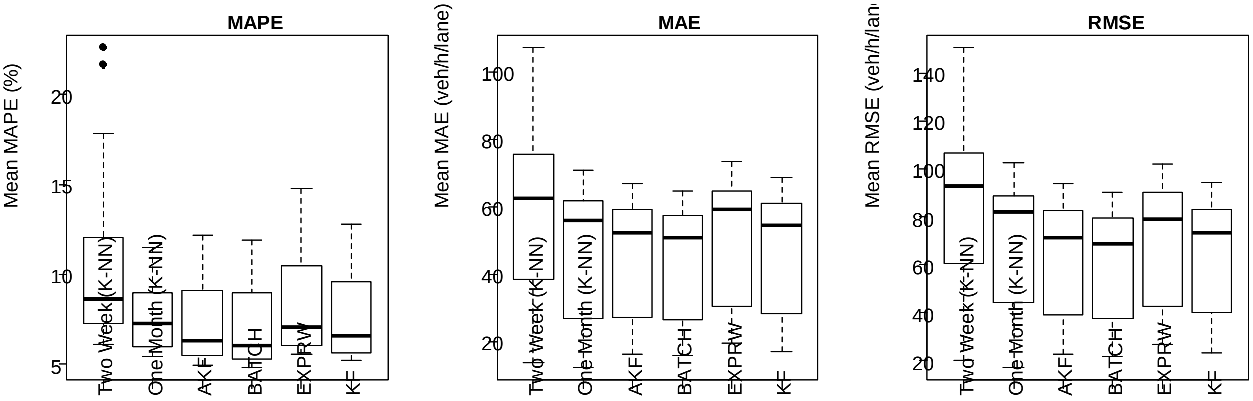

Figure 15 The impact of size of historic datasets (search space) on forecast accuracy

4.8. Comparing the enhanced and naïve K-NN methods

A number of enhancements, as described in this paper, are incorporated in the proposed K-NN based traffic forecasting method such as loess smoothing of the lagging parts of traffic profiles to reduce noise, weighted Euclidean distance to give more weight to recent measurements, winsorization of the candidate values to damps the effect of dominant candidates, and rank exponent method of aggregating the candidate values. It is important to quantify the improvements in forecast error obtained from using the proposed enhanced K-NN method as compared to naïve K-NN which uses Euclidean distance of the raw traffic flow rate profiles, simple arithmetic average of the candidate values without a mechanism to damp the effect of dominant candidates.

Figure 16 shows the forecast errors for the naïve K-NN method by level of traffic and time of the day. The box plots in Figure 16 show the spread of the forecast errors for naïve K-NN while the mean errors for the naïve K-NN and the enhanced K-NN methods are shown in solid red and broken green lines, respectively. Comparing the box plots of the forecast errors for the naïve K-NN (shown in Figure 16) and enhanced K-NN (shown in Figure 7), the distribution of the errors for the enhanced K-NN is narrower as the whisker error bars are less wide. Moreover, the results in Figure 16 indicate that there is a statistically significant increase in mean forecast errors between the proposed enhanced K-NN and naïve K-NN methods. The increases in the mean forecast errors from using the naïve K-NN are found to be 22% in MAPE, 25% in MAE, and 22% in RMSE. Enhanced K-NN has the best improvements during peak hours (or traffic level of Group 5) and its performance was more reliable than the naïve K-NN one. Note that for comparing the performance of the enhance K-NN with the naïve K-NN models, the entire datasets are used.

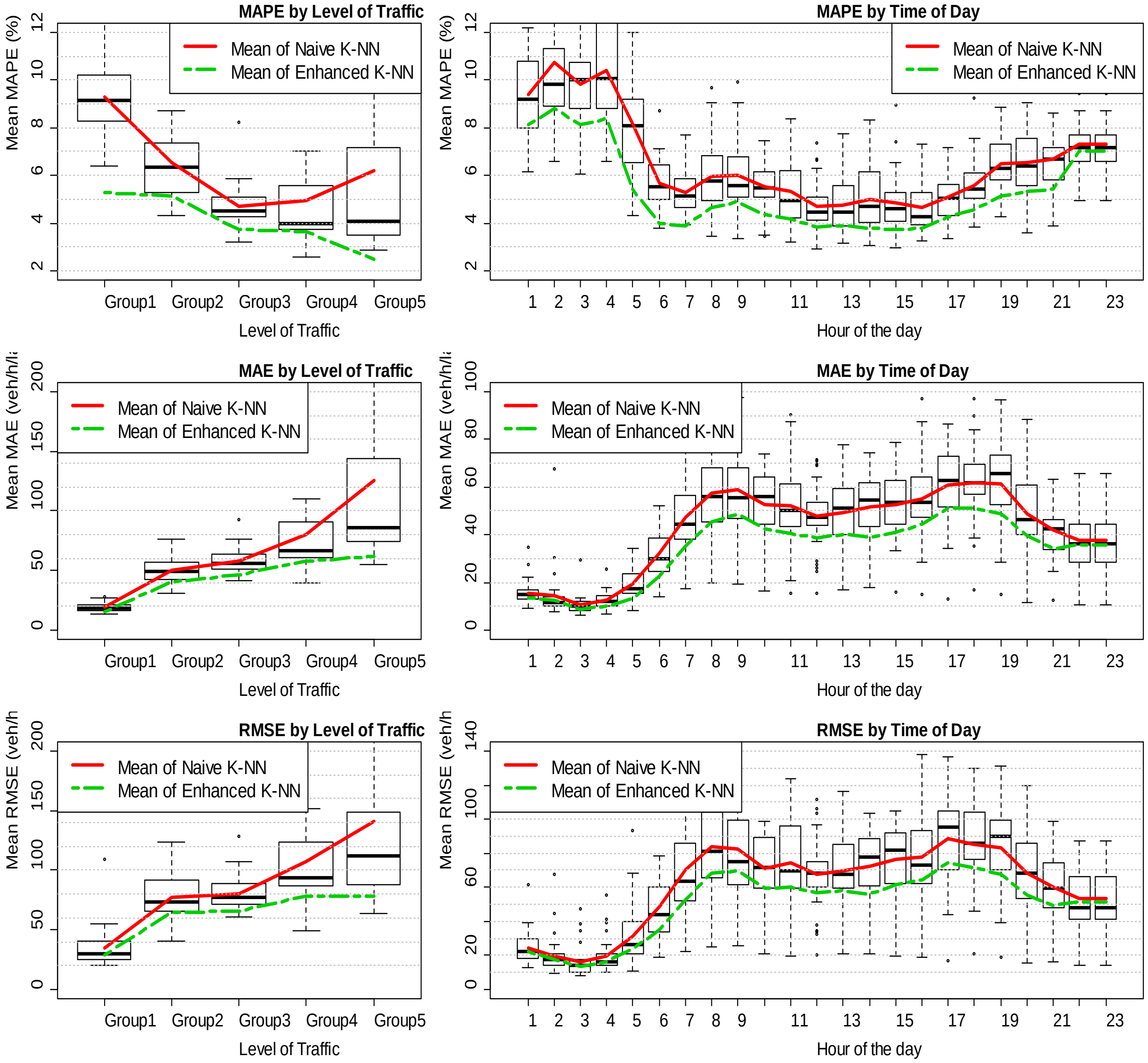

Figure 16 Forecast errors for naïve K-NN and its comparison with enhanced K-NN

5. Summary and Conclusions

The ability to timely and accurately forecast the evolution of traffic in the short future is very important for proactive traffic management strategies and provision of reliable travel times to travelers. In this paper, we presented a methodology for short-term traffic forecast which is solely data-driven employing the algorithm K-nearest neighbors using weighted Euclidean distance to identify similar traffic patterns. Moreover, variables of the K-NN algorithm were optimized, robustness of the proposed approach was demonstrated by applying it to large datasets collected from different regions, comparing it with other models, evaluating its performance for multiple steps, and testing the model under data with missing values. This research provides evidence that suggests the following key findings:

- The patterns of traffic that exist within the archived datasets can be used to provide reliable and accurate short-term traffic flow rate forecasts.
- Weighted Euclidean distance-based K-nearest neighbor algorithm is very effective in identifying similar traffic patterns from large sets of archived data.
- Non-parametric and data-driven approach for short-term traffic forecast provides better results when compared to models that use fitted equations, provided that enough data are available for selecting similar patterns.
- Given the simplicity, better accuracy, and robustness of the proposed approach, it can be easily incorporated with other traffic applications and real-time traffic control for proactive management of freeway traffic.
- Considering the accuracy of K-NN-based approach depends on size of the search space, models based on time series analysis should be used if the size of the available dataset is small.

One of the limitations of this study is that the proposed methodology was tested based on volume data collected at a single point on a freeway corridor, i.e., spatial features of traffic were not taken into consideration. However, this K-NN-based forecasting approach can be easily applied to multiple traffic stations spatially scattered throughout a road networks to obtain network level prediction. Another limitation is that the impact of factors that affect traffic operation, e.g., weather conditions, proportion of heavy vehicles, and events of incidents, etc., were not treated. Significant improvement in forecast accuracy may be obtained by giving proper attention to such events. The proposed enhanced K-NN approach can be used for general forecasting applications, e.g., speed, travel time, delays and other traffic variables. Future works will focus on prediction of traffic volumes at a network level as well as prediction of travel times on specified freeway corridors and incorporating exogenous factors that affect traffic operations into the forecast model.

Acknowledgements

The authors would like to thank Drs. Billy Williams and Jianhua Guo for providing the volume datasets used in this paper and responding to our inquiries about their published work on short-term forecasting. Funding for this research was in part provided by the Mid-Atlantic Transportation Sustainability University Transportation Center (MATS UTC).

6. References

Bajaw, S., Chung, E., and Kuwahara, M. (2003). A Travel Time Prediction Method Based on Pattern Matching Technique. Publ. ARRB Transp. Res. Ltd.

Castro-Neto, M., Jeong, Y.-S., Jeong, M.-K., and Han, L.D. (2009). Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions. Expert Syst. Appl. *36*, 6164–6173.

Cetin, M., and Comert, G. (2006). Short-term traffic flow prediction with regime switching models. Transp. Res. Rec. J. Transp. Res. Board *1965*, 23–31.

Chan, K.Y., Dillon, T.S., Singh, J., and Chang, E. (2012). Neural-network-based models for short-term traffic flow forecasting using a hybrid exponential smoothing and Levenberg–Marquardt algorithm. Intell. Transp. Syst. IEEE Trans. On *13*, 644–654.

Chen, H., and Rakha, H. (2014). Agent-Based Modeling Approach to Predict Experienced Travel Times. In Transportation Research Board 93rd Annual Meeting, (Washington D.C.),.

Chung, E., and Kuwahara, M. (2003). SENSITIVITY ANALYSIS OF SHORT-TERM TRAVEL TIME PREDICTION MODEL'S PARAMETERS.

Clark, S. (2003). Traffic prediction using multivariate nonparametric regression. J. Transp. Eng. *129*, 161–168.

Cleveland, W.S., and Devlin, S.J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. J. Am. Stat. Assoc. *83*, 596–610.

Cools, M., Moons, E., and Wets, G. (2009). Investigating the variability in daily traffic counts through use of ARIMAX and SARIMAX models. Transp. Res. Rec. J. Transp. Res. Board *2136*, 57–66.

Davis, G.A., and Nihan, N.L. (1991). Nonparametric regression and short-term freeway traffic forecasting. J. Transp. Eng. *117*, 178–188.

ELFAOUZI, N.-E. (1996). Nonparametric traffic flow prediction using kernel estimator. In International Symposium on Transportation and Traffic Theory, pp. 41–54.

Guo, J., Williams, B.M., and Smith, B.L. (2008). Data collection time intervals for stochastic short-term traffic flow forecasting. Transp. Res. Rec. J. Transp. Res. Board *2024*, 18–26.

Guo, J., Huang, W., and Williams, B.M. (2012). Integrated heteroscedasticity test for vehicular traffic condition series. J. Transp. Eng. *138*, 1161–1170.

Guo, J., Huang, W., and Williams, B.M. (2014). Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification. Transp. Res. Part C Emerg. Technol. *43*, 50–64.

Habtemichael, F., Cetin, M., and Anuar, K. (2015). Methodology for Quantifying Incident-Induced Delays on Freeways By Grouping Similar Traffic Patterns. In 94th Annual Meeting of the Transportation Research Board (TRB), (Washington D.C.),.

Hamed, M.M., Al-Masaeid, H.R., and Said, Z.M.B. (1995). Short-term prediction of traffic volume in urban arterials. J. Transp. Eng. *121*, 249–254.

Van Hinsbergen, J.W.C., and Sanders, F.M. (2007). Short Term Traffic Prediction Models. In 14th World Congress on Intelligent Transport System, (Beijing, China),.

Innamaa, S. (2005). Short-term prediction of travel time using neural networks on an interurban highway. Transportation *32*, 649–669.

Lee, S., and Fambro, D.B. (1999). Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting. Transp. Res. Rec. J. Transp. Res. Board *1678*, 179–188.

Li, C.-S., and Chen, M.-C. (2013). Identifying important variables for predicting travel time of freeway with non-recurrent congestion with neural networks. Neural Comput. Appl. *23*, 1611–1629.

Lin, L., Li, Y., and Sadek, A. (2013). A k Nearest Neighbor based Local Linear Wavelet Neural Network Model for On-line Short-term Traffic Volume Prediction. Procedia-Soc. Behav. Sci. *96*, 2066–2077.

Lin, L., Wang, Q., and Sadek, A. Short-Term Forecasting of Traffic Volume Evaluating Models Based on Multiple Data Sets and Data Diagnosis Measures. Transp. Res. Rec. *2392*, 40–47.

Van Lint, J., and Van Hinsbergen, C. (2012). Short term traffic and travel time prediction models, in artificial intelligence applications to critical transportation issues. In Transportation Research Circular, (Washington D.C.: National Academies Press),.

Lippi, M., Bertini, M., and Frasconi, P. (2013). Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. Intell. Transp. Syst. IEEE Trans. On *14*, 871–882.

Meade, N. (2002). A comparison of the accuracy of short term foreign exchange forecasting methods. Int. J. Forecast. *18*, 67–83.

Moorthy, C.K., and Ratcliffe, B.G. (1988). Short term traffic forecasting using time series methods. Transp. Plan. Technol. *12*, 45–56.

Mori, U., Medniburu, A., and Lozano, J. (2014). Distance Measures for Time Series data. User manual for Package TSdist.

Okutani, I., and Stephanedes, Y.J. (1984). Dynamic prediction of traffic volume through Kalman filtering theory. Transp. Res. Part B Methodol. *18*, 1–11.

Oswald, R.K., Scherer, W.T., and Smith, B.L. (2000). Traffic flow forecasting using approximate nearest neighbor nonparametric regression. Final Proj. ITS Cent. Proj. Traffic Forecast. Non-Parametr. Regres.

Al-Qahtani, F.H., and Crone, S.F. (2013). Multivariate k-nearest neighbour regression for time series data—A novel algorithm for forecasting UK electricity demand. In Neural Networks (IJCNN), The 2013 International Joint Conference on, (IEEE), pp. 1–8.

Rauscher, F.A. (1998). Learning to Predict the Duration of an Automobile Trip.

Robinson, S., and Polak, J. (2005). Modeling urban link travel time with inductive loop detector data by using the k-NN method. Transp. Res. Rec. J. Transp. Res. Board 47–56.

Ross, P. (1982). Exponential filtering of traffic data. Transp. Res. Rec. *869*, 43–49.

Smith, B.L., and Demetsky, M.J. (1997). Traffic flow forecasting: comparison of modeling approaches. J. Transp. Eng.

Smith, B.L., Williams, B.M., and Keith Oswald, R. (2002). Comparison of parametric and nonparametric models for traffic flow forecasting. Transp. Res. Part C Emerg. Technol. *10*, 303–321.

Szeto, W.Y., Ghosh, B., Basu, B., and O'Mahony, M. (2009). Multivariate traffic forecasting technique using cell transmission model and SARIMA model. J. Transp. Eng. *135*, 658–667.

Vlahogianni, E.I. (2007). Prediction of non-recurrent short-term traffic patterns using genetically optimized probabilistic neural networks. Oper. Res. *7*, 171–184.

Vlahogianni, E.I. (2008). Short-term predictability of traffic flow regimes in signalised arterials. Road Transp. Res. J. Aust. N. Z. Res. Pract. *17*, 19.

Vlahogianni, E.I., Golias, J.C., and Karlaftis, M.G. (2004). Short-term traffic forecasting: Overview of objectives and methods. Transp. Rev. *24*, 533–557.

Vlahogianni, E.I., Karlaftis, M.G., and Golias, J.C. (2014). Short-term traffic forecasting: Where we are and where we're going. Transp. Res. Part C Emerg. Technol.

Wang, J., and Shi, Q. (2013). Short-term traffic speed forecasting hybrid model based on Chaos–Wavelet Analysis-Support Vector Machine theory. Transp. Res. Part C Emerg. Technol. *27*, 219–232.

Wang, Y., and Papageorgiou, M. (2005). Real-time freeway traffic state estimation based on extended Kalman filter: a general approach. Transp. Res. Part B Methodol. *39*, 141–167.

Wang, J., Deng, W., and Guo, Y. (2014). New Bayesian combination method for short-term traffic flow forecasting. Transp. Res. Part C Emerg. Technol. *43*, 79–94.

Wang, Y., Papageorgiou, M., and Messmer, A. (2006). RENAISSANCE–A unified macroscopic model-based approach to real-time freeway network traffic surveillance. Transp. Res. Part C Emerg. Technol. *14*, 190–212.

Whittaker, J., Garside, S., and Lindveld, K. (1997). Tracking and predicting a network traffic process. Int. J. Forecast. *13*, 51–61.

Williams, B.M., and Hoel, L.A. (2003). Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. J. Transp. Eng. *129*, 664–672.

Xia, J., Huang, W., and Guo, J. (2012). A clustering approach to online freeway traffic state identification using ITS data. KSCE J. Civ. Eng. *16*, 426–432.

You, J., and Kim, T.J. (2000). Development and evaluation of a hybrid travel time forecasting model. Transp. Res. Part C Emerg. Technol. *8*, 231–256.

Zargari, S.A., Siabil, S.Z., Alavi, A.H., and Gandomi, A.H. (2012). A computational intelligence-based approach for short-term traffic flow prediction. Expert Syst. *29*, 124–142.

Zhang, L., Liu, Q., Yang, W., and Wei, N. (2013). An Improved K-nearest Neighbor Model for Short-term Traffic Flow Prediction. Procedia - Soc. Behav. Sci. *96*, 653–662.

Zheng, Z., and Su, D. (2014). Short-term traffic volume forecasting: A K-nearest neighbor approach enhanced by constrained linearly sewing principle component algorithm. Transp. Res. Part C Emerg. Technol. *43*, 143–157.

Zheng, W., Lee, D.-H., and Shi, Q. (2006). Short-term freeway traffic flow prediction: Bayesian combined neural network approach. J. Transp. Eng. *132*, 114–121.