Research Article

Evaluation of Short-Term Freeway Speed Prediction Based on Periodic Analysis Using Statistical Models and Machine Learning Models

Xiaoxue Yang,1 Yajie Zou,1 Jinjun Tang,2 Jian Liang,2 and Muhammad Ijaz1

1Key Laboratory of Road and Traffic Engineering of Ministry of Education, Tongji University, Shanghai 201804, China
2School of Traffic and Transportation Engineering, Smart Transport Key Laboratory of Hunan Province, Central South University, Changsha, China

Correspondence should be addressed to Yajie Zou; yajiezou@hotmail.com and Jinjun Tang; jinjuntang@csu.edu.cn

Received 7 June 2019; Revised 6 August 2019; Accepted 30 August 2019; Published 20 January 2020

Abstract

Accurate prediction of traffic information (i.e., traffic flow, travel time, traffic speed, etc.) is a key component of Intelligent Transportation System (ITS). Traffic speed is an important indicator to evaluate traffic efficiency. Up to date, although a few studies have considered the periodic feature in traffic prediction, very few studies comprehensively evaluate the impact of periodic component on statistical and machine learning prediction models. This paper selects several representative statistical models and machine learning approaches to analyze the influence of periodic component on short-term speed prediction under different scenarios: (1) multi-horizon ahead prediction (5, 15, 30, 60 minutes ahead predictions), (2) with and without periodic component, (3) two data aggregation levels (5-minute and 15-minute), (4) peak hours and off-peak hours. Specifically, three statistical models (i.e., space time (ST) model, vector autoregressive (VAR) model, autoregressive integrated moving average (ARIMA) model) and three machine learning approaches (i.e., support vector machines (SVM) model, multi-layer perceptron (MLP) model, recurrent neural network (RNN) model) are developed and examined. Furthermore, the periodic features of the speed data are considered via a hybrid prediction method, which assumes that the data consist of two components: a periodic component and a residual component. The periodic component is described by a trigonometric regression function, and the residual component is modeled by the statistical models or the machine learning approaches. The important conclusions can be summarized as follows: (1) the multi-step ahead prediction accuracy improves when considering the periodic component of speed data for both three statistical models and three machine learning models, especially in the peak hours; (2) considering the impact of periodic component for all models, the prediction performance improvement gradually becomes larger as the time step increases; (3) under the same prediction horizon, the prediction performance of all models for 15-minute speed data is generally better than that for 5-minute speed data. Overall, the findings in this paper suggest that the proposed hybrid prediction approach is effective for both statistical and machine learning models in short-term speed prediction.

1. Introduction

To alleviate the traffic congestion in large cities, it is particularly important to make full use of existing infrastructure resources such as the application of Intelligent Transportation System (ITS) [1–10]. Real-time and accurate prediction of traffic parameter, such as traffic flow, travel time, and travel speed, is an important input of ITS. Advanced Traveller Information System (ATIS) and Advanced Traffic Management System (ATMS) are essential parts of ITS, while dynamic traffic assignment (DTA) is a significant task for the operation of ATIS and ATMS. For the purpose of DTA, traffic flow and travel time were estimated and predicted to describe the traffic conditions in DynaMIT (Dynamic Network Assignment for the Management of Information to Travelers) system [10]. When traffic incidents occurred, the predicted travel time was used to evaluate the performance of the application of ITS based on DTA [11]. In particular, evacuation time was predicted to analyze the effectiveness of ITS for evacuation purpose [12]. Using the predicted information, road users can re-plan reasonable
travel modes and routes before traveling, and further adjust the route to improve travel efficiency. With rigorous structure and strong computational ability, a good prediction algorithm can usually capture all kinds of characteristics in traffic data. Up to date, lots of short-term traffic prediction methods have been proposed. Vlahogianni et al. [2, 3] provided a summary of existing short-term traffic methodological approaches until 2013 and divided them into two types: parametric methods and nonparametric methods. Van Lint and Van Hinsbergien [4] summarized the application of Neural Network in short-term traffic prediction. So far, traffic prediction models can be generally categorized into following several types: statistical methods, machine learning methods and hybrid methods.

Due to good theoretical interpretation ability and clear computational structure, statistical methods are widely applied to short-term traffic prediction. The conventional vector autoregressive (VAR) model and the autoregressive integrated moving average (ARIMA) model have been widely utilized in traffic prediction [13–16]. To improve the applicability of the conventional ARIMA model, Kumar and Vanajaski [17] developed a seasonal ARIMA (SARIMA) model. Zou et al. [16] used space time (ST) model to describe both time and space correlations of traffic. A spatio-temporal autoregressive moving average (STARIMA) model was proposed to utilize upstream volumes in the current moment to predict traffic condition [18]. Yang et al. [19] found that the spatial traffic information from upstream and downstream road segments can improve the prediction performance. Recently, some new models have also been developed. For example, Zhang et al. [20] proposed Granger causality to predict travel time and obtained better prediction performance. Agafonov and Yumaganov [21] presented a distributed model for short-term traffic flow prediction based on the k-nearest neighbors method.

In addition to statistical methods, the machine learning methods have been widely used in traffic prediction due to its strong generalization ability, learning ability and adaptability. Neural network is commonly used in traffic flow prediction [22]. Recurrent neural network (RNN) has been widely adopted for modeling nonlinear time series data because of its short term memory [23]. Recently, some researchers found that state-space neural network model [24] and long short-term memory neural network (LSTM) [25], which were improved based on the neural network, showed better prediction performance with high computational efficiency. Tang et al. [26] introduced an improved fuzzy-neural networks (FNN) to enhance traffic flow prediction accuracy. Dimitriou et al. [27] optimized the parameters of the adaptive fuzzy rule system using genetic algorithm. Lv et al. [28] proposed a deep-learning-based method using autoencoders to predict traffic flow. To capture the nonlinear spatial and temporal effect of traffic flow, Polson and Sokolov [29] introduced a deep learning approach combining a linear method. Based on mathematical methods and optimization techniques, support vector machine (SVM) transforms variables into a high dimensional space, and creates a hyperplane with maximum spacing. SVM contains two branches: support vector regression (SVR) and support vector classification (SVC). Furthermore, SVR model has a common application in traffic prediction [30, 31]. To speed up the parameter optimizing, the least square support vector machine (LS-SVMs) was used to predict traffic [32]. Jeong et al. [33] proposed an online learning weighted support-vector regression (OLWSVR) model to implement real-time traffic flow prediction. The Kaman filter theory (KFT), an algorithm for optimal estimation of system state using state equation of linear system, was used to predict traffic flow and travel time [34, 35]. Wang et al. [36] proposed an improved extended Kalman filter (EKF) for traffic time prediction. Guo et al. [37] introduced an adaptive Kalman filter (AKF) to predict unstable traffic flow.

Hybrid models combining the advantages of different methods are proposed to improve prediction accuracy. For example, the conditional probability theory and Bayesian rule were combined with ANN [38], and the statistical methods and heuristic models were combined with SVM [39]. Yanchong et al. [40] proposed a short-term traffic flow prediction model combining Mallat wavelet and BP neural network. Tang et al. [41] combined fuzzy c-means and Genetic algorithm to predict missing traffic volume data. Huang et al. [42] proposed a traffic flow prediction model based on fuzzy mean clustering (FCM) and advanced neural network (NN). To further improve the prediction performance, Huang et al. [43] developed a deep learning method incorporating a deep belief network (DBN) and a multilask regression layer. Tang et al. [44] introduced a hybrid model by combining double exponential smoothing (DES) and a support vector machine to predict traffic flow. It has been shown that hybrid models have better prediction performance than single method in traffic prediction.

Traffic data usually represent periodic features which help understanding the variation patterns of traffic flow and improving the prediction performance. To demonstrate the periodic pattern, researchers have introduced many prediction models considering cyclicity. To capture the weekly patterns of data, Williams and Hoel [45] applied Seasonal ARIMA model to traffic flow prediction. However, the outlier detection and the parameter estimation of Seasonal ARIMA model are time-consuming. Thus, Hong [46] applied seasonal support vector regression model with chaotic immune algorithm (SSVRCIA), Seasonal ARIMA model and seasonal Holt–Winters model to traffic flow prediction, and concluded that SSVRCIA model performed better than other models. Moreover, Lippi et al. [47] proposed two improved support vector regression models and compared the prediction performance in terms of accuracy and efficiency. Overall, Seasonal ARIMA model was proved to be more accurate; additionally, the new seasonal support vector regressor model performed better in peak hours. In addition, Li et al. [7] compared the prediction performance of hybrid models considering the periodicity of traffic time series data. And a frequently used strategy is the combination of prediction models and detrending methods [48–50]. For example, Dendrinos [51] divided the traffic-flow time series into two parts: periodic part and nonperiodic part, and took the nonperiodic part as the research focus. Some researchers modeled periodic components using spectral analysis technique [52, 53]. Zhang et al. [49] developed a hybrid traffic prediction method, which supposed that the traffic data contain three parts: periodic part, deterministic part, and volatility part. Furthermore, Zhang et al. [49] found that multi-step ahead prediction can provide more accurate prediction results.
Although a few studies have considered the periodic feature in traffic prediction, very few studies comprehensively evaluate the impact of periodic component on statistical methods and machine learning methods. Focusing on speed prediction, the specific research objectives are: (1) to evaluate the effectiveness of hybrid methods based on three statistical models (i.e., ST, VAR, ARIMA) and three machine learning models (i.e., SVM, Multi-layer Perceptron (MLP), RNN) in multi-step ahead prediction (5, 15, 30, 60 minutes ahead predictions) considering peak hours and off-peak hours, separately, (2) to compare the prediction performance improvement considering the impact of periodic component for all models, (3) to examine the difference in freeway speed prediction under two different data aggregation levels (5-minute and 15-minute).

The following sections of this paper are described below. The second part provides the data description, which shows the data gathered from five loop detectors of an eastbound road of Interstate 394 freeway stretch, Minnesota, from November 2017 to April 2018. The third part introduces two main methodologies used in this study: statistical models and machine learning approaches. The next section provides the modeling results. The short-term traffic speed prediction accuracy is evaluated under different scenarios. Finally, the model results are summarized and discussed.

2. Data Description and Preliminary Data Analysis

This study is carried out on an eastbound road of Interstate 394 freeway stretch, Minnesota, when suffering from heavy congestion during the rush hours of morning and afternoon. In this study, the road equipped with 5 neighboring stations is selected (see Figure 1). The length of selected road is about 1.7 miles. And the distance between two neighboring stations is approximately 0.5 miles. There are 3 lanes for the eastbound direction.

Speed data can be downloaded using the publicly available data tool developed by the Minnesota Department of Transportation. The speed data are collected for 5-minute and 15-minute aggregation level, respectively, 24 hours a day from loop detectors, from November 2017 to April 2018. The data missing rate is less than 0.01%, thus, the data repair method based on historical average is adopted to repair missing data. Specifically, traffic at night or on weekends is usually smooth and free of congestion. Therefore, this study selects the speed data from 6:00 AM to 8:00 PM of weekdays, which contain the morning and afternoon peak hours. Figure 1 provides the location of the selected I-394 road.

Figure 2 provides the distribution of median value of historical speeds on weekdays (Nov. 2017–April 2018) at all stations. From Figure 2, we can see that there are two peak hours for all stations: one is from about 7:00 AM to 9:00 AM and the other is from 3:30 PM to 7:30 PM.

It has been demonstrated that the traffic data of adjacent stations are spatially and temporally correlated [54, 55]. The speed values of stations in the downstream are influenced by the speed values of stations in the upstream. Furthermore, the traffic condition in the downstream also affects the traffic in the upstream because of traffic jam. In this study, cross-correlation function (CCF) is utilized to evaluate the time and spatial correlation of speed.

In this study, we choose station C as the target station. In Figure 3, the cross-correlation consequences between
station C and other adjacent stations are shown. As provided in Figure 3, absolute value of lag has a steady decreasing effect on the CCF values of speed. More specifically, the cross correlation value between station C and other stations is largest when the lag is equal to 0. Furthermore, when the absolute value of lag equals to 20, the CCF value is as low as 0.3. Based on Figure 3, as the distance between two stations increases, it can be observed that the correlation between two stations decreases, that is, the value presents a downward trend. As shown in Figure 3, 0.8543, 0.9342, 0.9000, and 0.7281 are the maximum CCF values between station C and stations A, B, D, and E, respectively. The result is reasonable, because the farther away the station is, the less impact it has.

Periodic patterns are another significant feature of traffic speed, except the temporal and spatial correlation above. In Figure 4, the speed distribution of station C during five continuous representative weekdays is shown. It has been shown that the period pattern exists in the traffic parameters [49, 56, 57]. As shown in Figure 4, the speed data show a periodic pattern every 24 hours. Obviously, during the peak hours in the morning and afternoon, the speed values suffer
from a significant decline. In the off-peak hours, the speed values suffer from the random fluctuation during the free flow traffic condition. Furthermore, similar periodic patterns have been demonstrated at four other stations. To capture this cyclical pattern, a hybrid prediction method is adopted in the following section.

3. Methodology

In this part, statistical models (i.e., ST, VAR, ARIMA), machine learning models (i.e., SVM, MLP, RNN) and hybrid models are described, respectively. In this study, due to its marked drop in speed over the peak hours, and the accessibility of speed data of loop detectors from the upstream and downstream, station C is chosen as the target station. The target station C is strongly correlated with the adjacent station, thus, the historical speed data of the adjacent station can be used to predict speed of the target station. In the section below, two aggregation levels (5-minute and 15-minute) are considered. Multi-horizon prediction (or 4-step ahead prediction under the aggregation level (for example, 5 minutes) is calculated by using the proposed methods, and the prediction horizons are 5 minutes, 15 minutes, and 1 hour.

3.1. Statistical Models. In this part, three statistical models are described as follows: ST model, VAR model, and ARIMA model.

3.1.1. ST Model. ST model, modeling via the proper probability distribution of data, is introduced to traffic flow and traffic speed prediction [16, 57]. The model contains two kinds of prediction values: point prediction and interval prediction. In this study, it is supposed that the speed at time \( t + p \) under the aggregation level \( i \) at station \( C \), \( C_{t+p,i} \), is assumed to follow a normal distribution [58], that is, \( C_{t+p,i} \sim N(\mu_{t+p,i}, \sigma^2_{t+p,i}) \).

\[ \mu_{t+p,i} = \alpha_0 + \alpha_1 C_{t,i} + \alpha_2 C_{t-1,i} + \alpha_3 B_{t,i} + \alpha_4 D_{t,i} + \alpha_5 D_{t-1,i} + \alpha_6 A_{t,i} + \alpha_7 A_{t-1,i} + \alpha_8 E_{t,i} + \alpha_9 E_{t-1,i}. \]  

(2)

Analyzing the speed data from November to December 2017, independent variables for \( \mu_{t+p,i} \) (for example, in (2)) are selected. The construction process of \( \mu_{t+p,i} \) is described as follows: starting from the simplest model, independent variables are added stepwise until Bayesian information criterion is not further improved [59].

\( \sigma_{t+p,i} \), the predictive spread, is fitted by a linear function of the fluctuating value \( v_{t,i} \):

\[ \sigma_{t+p,i} = b_0 + b_1 v_{t,i} \]  

(3)

where, coefficients \( b_0 \) and \( b_1 \) are nonnegative; and, the fluctuating value, which reflects the range of the recent fluctuations of speed, is fitted as follows:

\[ v_{t,i} = \left( \frac{1}{10} \sum_{j=0}^{4} \left( (A_{t,j} - A_{t-j,i})^2 + (B_{t,j} - B_{t-j,i})^2 + (C_{t,j} - C_{t-j,i})^2 + (D_{t,j} - D_{t-j,i})^2 + (E_{t,j} - E_{t-j,i})^2 \right)^{1/2} \right) \]

(4)

Figure 4: Periodic pattern of speed at station C during representative weekdays.
The range of the recent fluctuations of speed can be reflected by the fluctuating value.

3.1.2. VAR Model. Focusing on interrelated time series, VAR model can capture the effect of the upstream and downstream stations. In this study, a 5-equation VAR(m) model is defined as follows:

$$X_{t+1,j} = \Phi_0 + \Phi_1 X_{t,j} + \Phi_2 X_{t-1,j} + \cdots + \Phi_m X_{t-m+1,j} + u_{t+1,j},$$  \hspace{1cm} (5)

where, $X_{t+1,j} = (A_{t+1,j}, B_{t+1,j}, D_{t+1,j})^T$ is a 5 × 1 vector of variables; $\Phi_0$ is a 5 × 1 absolute term; $\Phi$ are matrices of coefficients; and, $u_{t+1}$ is a corresponding 5 × 1 independently and identically distributed random vector with $E(u_{t+1}) = 0$ and time invariant positive definite covariance matrix $E(u_{t+1}u_{t+1}^T) = \Sigma$.

Equation (5) can also be written as follows:

$$X_{t+1,j} = (A_{t+1,j} \Phi_1 + A_{t+1,1} \Phi_2 + \cdots + A_{t+1,m} \Phi_m) + (B_{t+1,j} \Phi_0 + B_{t+1,1} \Phi_1 + \cdots + B_{t+1,m} \Phi_m) X_{t,j} + (C_{t+1,j} \Phi_0 + C_{t+1,1} \Phi_1 + \cdots + C_{t+1,m} \Phi_m) X_{t-1,j} + (D_{t+1,j} \Phi_0 + D_{t+1,1} \Phi_1 + \cdots + D_{t+1,m} \Phi_m) X_{t-2,j} + \cdots + (E_{t+1,j} \Phi_0 + E_{t+1,1} \Phi_1 + \cdots + E_{t+1,m} \Phi_m) X_{t-m+1,j} + u_{t+1,j},$$  \hspace{1cm} (6)

3.1.3. ARIMA Model. Unlike VAR model, ARIMA model only takes the effect of time series into account, and has been used in various traffic data analysis [60–62]. A nonseasonal ARIMA model can capture the effect of the upstream and downstream stations. In this study, a 5-equation VAR(m) model is defined as follows: ARIMA (p, d, q) can be defined as follows: ARIMA ($\Phi$, $\Theta$, $\Psi$) is a Gaussian white noise series whose mean is zero and variance is $\sigma^2$; and, $\Phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ and $\Theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$ are polynomials of $\phi$ and $\theta$, and $\phi(z) \neq 0$ for $|z| \leq 1$. $\Phi$ to $\Psi$ are matrices of coefficients; and, $u_{t+1}$ is a corresponding 5 × 1 independently and identically distributed random vector with $E(u_{t+1}) = 0$ and time invariant positive definite covariance matrix $E(u_{t+1}u_{t+1}^T) = \Sigma$.

$$X_{t+1,j} = \Phi_0 + \Phi_1 X_{t,j} + \Phi_2 X_{t-1,j} + \cdots + \Phi_m X_{t-m+1,j} + u_{t+1,j},$$  \hspace{1cm} (8)

where, $X_{t+1,j} = (A_{t+1,j}, B_{t+1,j}, D_{t+1,j})^T$ is a 5 × 1 vector of variables; $\Phi_0$ is a 5 × 1 absolute term; $\Phi$ to $\Psi$ are matrices of coefficients; and, $u_{t+1}$ is a corresponding 5 × 1 independently and identically distributed random vector with $E(u_{t+1}) = 0$ and time invariant positive definite covariance matrix $E(u_{t+1}u_{t+1}^T) = \Sigma$.

Through the evaluation of characteristic polynomial, the stabilization of VAR(m) model can be guaranteed:

$$\det(I_{5} - \Phi_1 z - \cdots - \Phi_m z^m) \neq 0 \text{ for } |z| \leq 1.$$  \hspace{1cm} (7)

3.2. Machine Learning Models. In this section, we describe three machine learning models: SVM model, MLP model, and RNN model.

3.2.1. SVM Model. SVM transforms vector variables into a higher dimensional space, and creates a hyperplane with maximum spacing. SVM contains two branches: support vector regression (SVR) and support vector classification (SVC). SVC, addressing the classification problems, calculates a decision boundary and maximizes the distance between the boundary and the nearest sample data. Like SVC, SVR uses a similar approach for regression problems and ignores the error which is less than $\epsilon$ between the observed value and the estimated value [60, 63, 64]. More specifically, given a group of training data, the objective is to seek a function $f(x)$ that the maximum deviation between actual values $y_i$ and predicted values is at most $\epsilon$. For instance, a linear function $f(x) = w^T x + b$ is flat if it has a small $\omega$—this can be achieved by minimizing $||\omega||^2$. Due to the function that satisfies all the necessary constraints $\mathcal{C}$ may not exist, some slack variables $(\xi, \xi^*)$ are introduced to allow for some errors.

The formulation of SVR is defined as follows:

$$\minimize \left\{ \frac{1}{2} ||\omega||^2 + C \sum_{i=1}^{n} (\xi + \xi^*) \right\}$$

subject to

$$y_i - \omega^T x_i - b \leq \epsilon + \xi_i,$$

$$\omega^T x_i + b - y_i \leq \epsilon + \xi_i^*,$$

$$\xi, \xi^* \geq 0.$$  \hspace{1cm} (10)

where, $f(x) = \omega^T x + b$ is the objective function; $y_i$ is actual value; $\epsilon$ is the deviation between actual values and predicted values; and, $\xi, \xi^*$ are slack variables allowing for some errors.

We can also extend SVR to nonlinear problems by combining nonlinear kernel functions. Common kernels include linear kernel and the Radial Basis Function (RBF) kernel, which convert the input sample into a higher dimensional space that results in better separation (for classification) or estimation of values (for regression). In this study, we experimentally choose to use a RBF kernel for SVR because it generally provides better results.
3.2.2. **MLP Model.** Multilayer perception is the most common Artificial Neural Network (ANN) model[61, 65, 66]. Jiang et al. [67] selected MLP model as one of the candidates to compare the prediction accuracy of different models. A neural network consists of some layers, each of which has one or more neurons. Every neuron is linked with all the neurons in the adjacent layer, while the neurons in the layer are not. Each neuron takes a linear weighted combination of all its input \( x \) (from the layer in front of it) and generates output \( y \) through a nonlinear activation function:

\[
y = \sigma \left( \sum_{i=1}^{N} (\omega_i x_i) \right).
\]  

where, \( \sigma \) is the hidden layer function and is usually a sigmoid function.

Each of these outputs \( y \) is used as an input to the next layer of neurons until the last (i.e., output) layer is reached. The weights correlated with each neuron can be randomly initialized so that each neuron can potentially learn a different function of its input.

And the loss function measures the difference between the estimated output value of the network and the real value of the training data. For regression problems, more specifically, the squared error between the estimate and the actual value is often used as a loss function. The backpropagation algorithm is then utilized to compute the gradient of this error, and the gradient is propagated back through the network (towards the input layer), and the weight of each neuron is updated by gradient descent. The weights associated with each neuron are the parameters that define the neural network model, which are estimated by minimizing a loss function. A random gradient-based optimizer called Adam, which is computationally efficient and extends well to larger data sets, is adopted [68]. Set all parameters of this optimizer to their default values. The rectified linear unit (ReLU) activation function is used in the MLP network.

3.2.3. **RNN Model.** As shown in Figure 5, RNN is a particular neural network, which consists of at least one feedback link that serves as the internal state from the neuron’s outputs to neural network, which consists of at least one feedback link.

\[
y_t = H(Ux_t + Ws_{t-1} + b),
\]

\[
o_t = Vs_t + b_o,
\]

where, \( x = (x_1, ..., x_T) \) is the input vector; \( s = (s_1, ..., s_T) \) is the hidden vector; \( o = (o_1, ..., o_T) \) is the output vector; \( W, U, \) and \( V \) are the weight matrices; \( b_s \) and \( b_o \) are the bias vectors; and, \( H \) represents the hidden layer function and is usually a sigmoid function.

3.3. **Hybrid Models.** As shown in Figure 4, the traffic speed usually demonstrates a daily cyclical pattern. Therefore, it is feasible to decompose the speed data into two components: periodic component which demonstrates the cyclical trend in the weekdays, and a residual component. Therefore, the hybrid prediction methods can be defined as follows:

\[
C_t = M_t + C_t',
\]

where, \( C_t \) is the original speed at station \( C_t \); \( M_t \) is the cyclical trend at station \( C_t \); and, \( C_t' \) is the residual component at station \( C_t \).

A trigonometric regression function, combining sinusoids and cosinoids, is used to fit the cyclical pattern observed in Figure 4. Adorf [72] applied the trigonometric regression function to capture the periodic patterns when analyzing time series data. Furthermore, to improve the prediction performance for the wind speed, Gneiting et al. [59] adopted the trigonometric functions in the ST models and achieved accurate prediction results.

Taking the aggregation level of 5-minute as an example, the average speed of every station is computed as follows:

\[
S_t = \frac{1}{30} \sum_{d=1}^{30} s^d_t,
\]

where, \( S_t \) is the average speed at time \( t \); \( s^d_t \) is 5-minute average speed at time \( t \) on day \( d; t = 1, 2, ..., 288; \) and, \( d = 1, 2, ..., 30 \) is the number of days.

The periodic component can be fitted as follows:

\[
M_u = m_0 + m_1 \sin \left( \frac{2\pi u}{288} \right) + m_2 \cos \left( \frac{2\pi u}{288} \right) + m_3 \sin \left( \frac{4\pi u}{288} \right) + m_4 \cos \left( \frac{4\pi u}{288} \right) + \ldots + m_{2n-1} \sin \left( \frac{2\pi u}{288} \right) + m_{2n} \cos \left( \frac{2\pi u}{288} \right),
\]

where, \( u = 1, 2, ..., 288; n \) is the number of trigonometric polynomials.

Previous study [16] has examined different number of trigonometric functions on speed prediction and found that the prediction accuracy improves slightly after \( n \) reaches 15. Thus, we set \( n \) to 15 in this study.

The residual component is fitted by the ST, VAR, ARIMA, SVM, MLP, and RNN models. Statistical prediction models (i.e., ST, HST (hybrid ST), VAR, HVAR (hybrid VAR), ARIMA, and HARIMA (hybrid ARIMA)) are estimated in the Software R, and machine learning models (i.e., SVM, H SVM (hybrid SVM), MLP, HMLP (hybrid MLP), RNN, and HRNN (hybrid RNN)) are estimated in the Software Python.
4. Results

4.1. Evaluation Indicators. To calculate the multi-step ahead prediction performance of different models, three performance evaluation indexes (i.e., the mean absolute error (MAE), the mean absolute percentage error (MAPE), and the root mean square error (RMSE)) are adopted. The calculation formulas of these three indicators are shown as follows:

\[
\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |s'_i - s_i|, \tag{17}
\]

\[
\text{MAPE} = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{|s'_i - s_i|}{s_i} \right) \times 100\%, \tag{18}
\]

\[
\text{RMSE} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (s'_i - s_i)^2}. \tag{19}
\]

where, \(N\) is the number of observations; \(s_i\) is the actual speed at time \(i\) at station \(C\); and, \(s'_i\) is the predicted speed at time \(i\) at station \(C\).

To evaluate the performance of all models, both one-step and multi-step ahead prediction are considered.

4.2. Comparison of Prediction Results. In this study, the prediction performance of ST, VAR, ARIMA, SVM, MLP, RNN, and hybrid models is compared using the speed data at station \(C\). The data samples collected during 6:00 AM–8:00 PM from April 2nd to April 30th (21 weekdays) are selected as the testing period. Specifically, we utilize a sliding training period that consists of the last 65 days before the prediction point for ST, VAR, and ARIMA models. For instance, the parameters of models for speed prediction on April 10, 2018 are calculated using the last 65 days (i.e., April (6 days), March (22 days), February (20 days), and January (17 days)). To be fair, the data samples from January 1st to March 30th (65 weekdays) are chosen as the training period to optimize parameters in machine learning models.

In the prediction, the optimal parameters of statistical models are calculated in \(\hat{\theta}\). More specifically, ARIMA and VAR models use forecast and vars packages. The optimal order in ARIMA model is chosen by the Akaike Information Criterion (AIC) values with the last 65 days of data. A maximal order of 10 is used in VAR model. And the optimal order of VAR model is also determined via the AIC values. All machine learning algorithms are implemented using scikit-learn and keras software packages in the Python programming language. Radial Basis Function (RBF) is set as the kernel functions in SVM. All models consider multi-horizon ahead prediction (5, 15, 30, 60 minutes ahead predictions).

From Figure 2, we can see that there are two peak hours for all stations: one is about from 7:00 AM to 9:00 AM and the other is from 3:30 PM to 7:30 PM. Thus, we divide the target period into two parts: peak hours and off-peak hours.

Figure 6 shows prediction results of MAE values of from 5 minutes to 60 minutes ahead prediction over all models with different aggregation levels in peak hours. In terms of 5 minutes ahead prediction, the performance of hybrid models is slightly better than that of single models. The average MAE
value of HARIMA model is 4.7999 while the value of ARIMA model is 4.7683 for 5 minutes ahead prediction. Specifically, the improvement of prediction accuracy is 0.66%, that is, both the hybrid model and the single model perform equally well for 5 minutes ahead prediction. However, for 5-minute data, the MAE value for HMLP model is 11.5593 and the value for MLP model is 15.1838 in terms of 60 minutes ahead prediction. That demonstrates the improvement is 23.87% in prediction performance. Overall, hybrid models perform better than single models in terms of MAE values. Hybrid models can provide explicit description of the basic structure of data and give better insights into potential characteristics of time series data. In addition, Figures 7 and 8 summarize the results of MAPE values and RMSE values. Moreover, similar findings with MAE values can be drawn: for 15-minute speed data, hybrid models show slight improvement in terms of MAPE values and RMSE values of one-step ahead prediction. Moreover, for 60 minutes ahead prediction, hybrid model improves accuracy by 38.43% of MAPE value for ARIMA model, and 26.46% is the improvement of the RMSE value for SVM model. In summary, hybrid models outperform the ST, VAR, ARIMA, SVM, MLP, and RNN models in terms of prediction accuracy. This is because the hybrid models, considering the trigonometric polynomials, can better capture the periodic features of speed data; (2) under the same prediction horizon, the prediction performance of hybrid models for 15-minute speed data is generally better than that for 5-minute speed data.

In summary, two interesting findings can be summarized as follows: (1) the hybrid models can outperform the ST, VAR, ARIMA, SVM, MLP, and RNN models in terms of the prediction accuracy. This is because hybrid models, considering the trigonometric polynomials, can better capture the periodic features of speed data; (2) under the same prediction horizon, the prediction performance of hybrid models for 15-minute speed data is generally better than that for 5-minute speed data.

Figures 9, 10, and 11 provide MAE, MAPE, and RMSE values of different prediction horizons for all models in off-peak hours. Compared with peak-hours, the prediction performance in off-peak hours is better on both single models and hybrid models. For example, for 5-minute speed data, the MAE value for HVAR model of 30 minutes ahead prediction in peak hours, 10.4796, is much larger than that in off-peak hours, 4.4149. In addition, similar conclusions can be found in terms of MAPE values and RMSE values. The possible reason is described as follows: the speed in off-peak hours is close to free-flow speed and fluctuates slightly. To be specific, the
is selected to construct the periodic component of daily similarity, while three statistical models (space time model, ARIMA model, and VAR model) and three machine learning models (SVM model, MLP model, and RNN model) are used to describe the residual component. Although statistical models and machine learning models both aim to predict traffic speed, these two types of approaches differ greatly in model structure and result interpretation. Statistical models, based on a rule-based mechanism, are presented in mathematical formula form. Additionally, each variable in the formula has a specific or practical meaning. However, the application of statistical models is generally limited to the assumption embedded in the model. For example, data are commonly required to vary linearly with time in ARIMA model and VAR model. In contrast, machine learning models directly learn from data without any explicit model structure, and operate as a “black box.” Thus, the application of these models is more flexible, especially when handling big data. Nevertheless, lacking the capability of result interpretation is generally the limitation of these models. Given the significant difference between these two types of prediction, this study explored whether periodic components have the same impact on statistical model or machine learning model. Consequently, based on the above results, it can be inferred that the periodic component improves prediction performance both for statistical model and machine learning model.

5. Discussion

5.1. Effects of Hybrid Models. Decomposing original data into two components (residual component and periodic component), hybrid models can capture the dynamic fluctuations of traffic data accurately. The periodic component, the main trend of time series, generally represents the time-independent part of original data; while the residual component reflects the time-dependent part of time series. In this study, a trigonometric regression function is selected to construct the periodic component of daily similarity, while three statistical models (space time model, ARIMA model, and VAR model) and three machine learning models (SVM model, MLP model, and RNN model) are used to describe the residual component. Although statistical models and machine learning models both aim to predict traffic speed, these two types of approaches differ greatly in model structure and result interpretation. Statistical models, based on a rule-based mechanism, are presented in mathematical formula form. Additionally, each variable in the formula has a specific or practical meaning. However, the application of statistical models is generally limited to the assumption embedded in the model. For example, data are commonly required to vary linearly with time in ARIMA model and VAR model. In the contrast, machine learning models directly learn from data without any explicit model structure, and operate as a “black box”. Thus, the application of these models is more flexible, especially when handling big data. Nevertheless, lacking the capability of result interpretation is generally the limitation of these models. Given the significant difference between these two types of prediction, this study explored whether periodic components have the same impact on statistical model or machine learning model. Consequently, based on the above results, it can be inferred that the periodic component improves prediction performance both for statistical model and machine learning model.
Specifically, the impact of periodic component in peak hours on both statistical model and machine learning model is examined. From one-step ahead prediction to 12-step ahead prediction, the improvement of accuracy ranges from 0.65% to 28.37% of ARIMA model in terms of 5-minute data. Similar conclusions can be drawn from other models (VAR, HVAR, HARIMA, HSVM, HMLP, HRNN).

5.2. Improvements of Hybrid Models over Single Models. Improvement percentage, the ratio of prediction accuracy improvement from single models to hybrid models, is utilized to measure the effect considering periodic component. Compared with off-peak hours, the prediction performance in peak hours is generally more critical.
predictions are required, then it is worthy dedicating further study into effectively utilizing high-aggregation data while minimizing the negative effects of noise. Additionally, in terms of low-aggregation data, it is required to reduce information loss while ensuring high prediction accuracy.

6. Conclusions

In this study, we examine the impact of periodic component on three statistical models and three machine learning models when predicting freeway traffic speed using the data collected from five loop detectors of an eastbound road of Interstate 394 freeway stretch, Minnesota. In addition, multi-horizon ahead prediction and two aggregation levels of data are also considered. The prediction performance is measured by three evaluation indicators: MAE, RMSE, and MAPE. The important conclusions can be summarized as follows: (1) the multi-step prediction accuracy improves considering the periodic component of speed data for both statistical models and machine learning models, especially in the peak hours; (2) as the time step increases, the prediction performance improvement gradually becomes larger considering the impact of periodic component over all models; (3) under the same prediction horizon, the prediction performance of all models for 15-minute speed data is generally better than that for 5-minute speed data. Thus, the selection of prediction models can significantly affect the prediction accuracy in terms of MAE, MAPE and RMSE, and further influences the performance regarding traffic management tools such as ATIS and ATMS of ITS system. Inaccurate prediction of traffic parameters may have a significant impact...
on the operation of ITS system, which aims to improve the efficiency and capacity of traffic and transportation system. Given the accurate real-time information about traffic condition (traffic parameters such as traffic speed, traffic flow, and travel time), drivers can replan or adjust travel routes to avoid the heavy traffic of original routes. Additionally, either ramp controlling or variable speed limiting can effectively alleviate traffic congestion while the traffic speed of mainline closes to the critical speed. In summary, transportation management agencies should be cautious further when predicting traffic condition to maintain the regular operation of ITS [45].

In the future, several in-depth research can be conducted. First, nonrecurring congestion events, such as incidents and special events, will influence the distribution of traffic speed. Thus, it is necessary to examine the prediction performance of these models for nonrecurring congestion condition. Second, to address the heterogeneous speed data, some prediction methods based on finite mixture model and copula model may be developed [73, 74]. Third, it is useful to compare the traffic prediction performance between the traffic flow models [75–77] and the proposed prediction models in this study. Fourth, the possible impact of periodic component on interval prediction should also be considered [37, 78–80].

Data Availability

The data used to support the findings of this study are available from the publicly available data tool developed by the Minnesota Department of Transportation. Please visit: http://data.dot.state.mn.us/datatools/.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was funded jointly by the National Key Research and Development Program of China, grant number 2018YFE0102800; the National Natural Science Foundation of China, grant number 71971160, 71701215; Shanghai Science and Technology Committee, grant number 18510745400; Foundation of Central South University, grant number 502045002; Science and Innovation Foundation of the Transportation Department in Hunan Province, grant number 201725; Postdoctoral Science Foundation of China, grant number 140050005.

Table 1: Improvement percentage of hybrid models for different prediction horizons in peak hours at station C (time scale: 5 min).

<table>
<thead>
<tr>
<th>MAE</th>
<th>Prediction horizons (minutes)</th>
<th>5</th>
<th>15</th>
<th>30</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST</td>
<td>0.1678</td>
<td>2.0201</td>
<td>6.5398</td>
<td>18.4136</td>
<td></td>
</tr>
<tr>
<td>VAR</td>
<td>0.7073</td>
<td>2.4666</td>
<td>7.7073</td>
<td>16.9290</td>
<td></td>
</tr>
<tr>
<td>ARIMA</td>
<td>0.6583</td>
<td>2.7986</td>
<td>13.5338</td>
<td>28.3774</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>2.7703</td>
<td>0.8935</td>
<td>9.5560</td>
<td>21.5672</td>
<td></td>
</tr>
<tr>
<td>MLP</td>
<td>0.3708</td>
<td>0.6867</td>
<td>5.3738</td>
<td>23.8708</td>
<td></td>
</tr>
<tr>
<td>RNN</td>
<td>1.4256</td>
<td>4.5202</td>
<td>8.9773</td>
<td>15.5103</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MAPE (%)</th>
<th>Prediction horizons (minutes)</th>
<th>5</th>
<th>15</th>
<th>30</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST</td>
<td>−0.0816</td>
<td>3.1284</td>
<td>11.1746</td>
<td>27.1694</td>
<td></td>
</tr>
<tr>
<td>VAR</td>
<td>5.5966</td>
<td>12.2921</td>
<td>17.4324</td>
<td>27.7299</td>
<td></td>
</tr>
<tr>
<td>ARIMA</td>
<td>3.1773</td>
<td>10.5383</td>
<td>23.3195</td>
<td>38.2678</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>5.0060</td>
<td>2.4855</td>
<td>16.7698</td>
<td>35.2814</td>
<td></td>
</tr>
<tr>
<td>MLP</td>
<td>2.4710</td>
<td>5.2158</td>
<td>10.1209</td>
<td>38.1002</td>
<td></td>
</tr>
<tr>
<td>RNN</td>
<td>−1.3680</td>
<td>4.7849</td>
<td>18.4671</td>
<td>25.7139</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RMSE</th>
<th>Prediction horizons (minutes)</th>
<th>5</th>
<th>15</th>
<th>30</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST</td>
<td>2.2066</td>
<td>7.9375</td>
<td>12.1864</td>
<td>21.1225</td>
<td></td>
</tr>
<tr>
<td>VAR</td>
<td>1.4003</td>
<td>−2.3589</td>
<td>4.3848</td>
<td>17.1806</td>
<td></td>
</tr>
<tr>
<td>ARIMA</td>
<td>5.0394</td>
<td>11.0708</td>
<td>20.5272</td>
<td>32.1107</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>5.2385</td>
<td>7.5790</td>
<td>16.1069</td>
<td>28.5130</td>
<td></td>
</tr>
<tr>
<td>MLP</td>
<td>3.0359</td>
<td>9.1199</td>
<td>15.6351</td>
<td>28.6400</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Improvement percentage of hybrid models for different prediction horizons in peak hours at station C (time scale: 15 min).

<table>
<thead>
<tr>
<th>MAE</th>
<th>Prediction horizons (minutes)</th>
<th>5</th>
<th>15</th>
<th>30</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST</td>
<td>—</td>
<td>2.8043</td>
<td>9.6993</td>
<td>19.3439</td>
<td></td>
</tr>
<tr>
<td>VAR</td>
<td>—</td>
<td>4.6387</td>
<td>9.3649</td>
<td>18.4457</td>
<td></td>
</tr>
<tr>
<td>ARIMA</td>
<td>—</td>
<td>3.3953</td>
<td>10.6782</td>
<td>26.6451</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>—</td>
<td>4.7455</td>
<td>10.8446</td>
<td>18.7158</td>
<td></td>
</tr>
<tr>
<td>MLP</td>
<td>—</td>
<td>2.4939</td>
<td>9.0440</td>
<td>6.4904</td>
<td></td>
</tr>
<tr>
<td>RNN</td>
<td>—</td>
<td>11.3976</td>
<td>16.7804</td>
<td>18.3007</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MAPE (%)</th>
<th>Prediction horizons (minutes)</th>
<th>5</th>
<th>15</th>
<th>30</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST</td>
<td>—</td>
<td>3.3989</td>
<td>12.8081</td>
<td>28.2290</td>
<td></td>
</tr>
<tr>
<td>VAR</td>
<td>—</td>
<td>12.6058</td>
<td>20.0241</td>
<td>29.9824</td>
<td></td>
</tr>
<tr>
<td>ARIMA</td>
<td>—</td>
<td>10.6009</td>
<td>21.7511</td>
<td>38.4383</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>—</td>
<td>10.4714</td>
<td>22.4997</td>
<td>30.4860</td>
<td></td>
</tr>
<tr>
<td>MLP</td>
<td>—</td>
<td>5.8083</td>
<td>18.3380</td>
<td>14.1194</td>
<td></td>
</tr>
<tr>
<td>RNN</td>
<td>—</td>
<td>9.1287</td>
<td>14.0741</td>
<td>28.1865</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RMSE</th>
<th>Prediction horizons (minutes)</th>
<th>5</th>
<th>15</th>
<th>30</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST</td>
<td>—</td>
<td>7.6198</td>
<td>14.7111</td>
<td>21.9747</td>
<td></td>
</tr>
<tr>
<td>VAR</td>
<td>—</td>
<td>6.7191</td>
<td>9.1703</td>
<td>18.1812</td>
<td></td>
</tr>
<tr>
<td>ARIMA</td>
<td>—</td>
<td>10.0176</td>
<td>18.3135</td>
<td>31.1368</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>—</td>
<td>8.9609</td>
<td>15.2175</td>
<td>26.4659</td>
<td></td>
</tr>
<tr>
<td>MLP</td>
<td>—</td>
<td>9.7862</td>
<td>18.0127</td>
<td>11.4506</td>
<td></td>
</tr>
<tr>
<td>RNN</td>
<td>—</td>
<td>12.2067</td>
<td>19.4621</td>
<td>23.5209</td>
<td></td>
</tr>
</tbody>
</table>

References


