Multi-Step-Ahead Traffic Speed Forecasting Using Multi-Output Gradient Boosting Regression Tree

Xingbin Zhan[a], Shuaichao Zhang[b], Wai Yuen Szeto[a,c], Xiqun (Michael) Chen[b,*]

[a] *Department of Civil Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong, China*
[b] *College of Civil Engineering and Architecture, Zhejiang University, Hangzhou 310058, China*
[c] *The University of Hong Kong Shenzhen Institute of Research and Innovation, Shenzhen, China*

[*] Corresponding author. Email: chenxiqun@zju.edu.cn Tel.: (+86)571-88208938
Mailing address: B828 Anzhong Building, Zhejiang University, 866 Yuhangtang Rd, Hangzhou 310058, China

Abstract

Short-term traffic speed forecasting is an important component of Intelligent Transportation Systems (ITS). Multi-step-ahead prediction can provide more information and predict the longer trend of traffic speed than single-step-ahead prediction. This paper presents a multi-step-ahead traffic speed prediction approach by improving the gradient boosting regression tree (GBRT). The traditional multiple output strategies, e.g., the direct strategy and iterated strategy, share a common feature that they model the samples through multi-input single-output mapping rather than multi-input multi-output mapping. This paper proposes multivariate GBRT to realize simultaneous multiple outputs by considering correlations of the outputs which have not been fully considered in the existing strategies. For illustrative purposes, traffic detection data are extracted at the 5-min aggregation time interval from three loop detectors in US101-N freeway through the Performance Measurement System (PeMS). The support vector regression (SVR) is used as the benchmark. Assessments on the three models are based on the three criteria, i.e., prediction accuracy, prediction stability, and prediction time. The results indicate that (I) Multivariate GBRT and GBRT using the direct strategy have higher prediction accuracies compared with SVR; (II) GBRT using the iterated strategy has a good prediction accuracy in short-step-ahead prediction and the prediction accuracy decreases significantly in long-step-ahead prediction; (III) Multivariate GBRT has the best stability which means the higher reliability in multi-step-ahead prediction while iterated GBRT has the worst stability; and (IV) Multivariate GBRT has an enormous advantage in the prediction efficiency and this advantage will expand with the increasing prediction horizons.

Keywords: Traffic speed forecasting; multivariate GBRT; multi-step-ahead prediction; direct strategy; iterated strategy

1. INTRODUCTION

With the development of social economy and motorization, traffic congestion of urban road networks has been more and more serious. The realization of short-term traffic speed forecasting is the prerequisite and key of many intelligent transportation systems (ITS), e.g., traffic guidance system, traffic signal optimization system, and vehicle scheduling management system. Compared with the one-step-ahead forecasting of traffic speed, the multi-step-ahead prediction of short-term traffic speed is more necessary and meaningful in practice. For traffic policymakers and regulators, the multi-step-ahead prediction information is conducive to the development of more effective traffic control strategies to alleviate traffic congestion, which improves the allocation efficiency of human and material resources. For informed travelers, the multi-step-ahead forecasting information is beneficial for pre-trip route planning or en-route decisions. However, due to the fluctuation and uncertainty of traffic speed, the multi-step-ahead prediction can be easily disturbed by many random factors, e.g., travel demand fluctuations, weather conditions, accidents, and road work, which make the multi-step-ahead prediction more difficult than the one-step-ahead prediction.

As one of the fundamental parameters to evaluate traffic states, traffic speed can be easily understood and accepted by both travelers and managers. Based on the predicted speed, informed travelers can select the route, departure time, and travel mode more effectively. Traffic speed prediction is a complicated and challenging task. It usually fluctuates strongly at different times due to the interference of external factors such as driver behavior, weather, incidents, and road surface conditions. These fluctuations are usually non-linear and complex. So it is important to fully understand these fluctuations and develop accurate prediction algorithms for short-term forecasting. The measured speed data of the target observation location and its adjacent locations are spatially correlated, meanwhile, the speeds measured at different time intervals are temporally correlated. Considering both the temporal and spatial correlations will have a positive effect on the forecasting accuracy.

This paper improves the gradient boosting regression tree (GBRT) to achieve the multi-step-ahead prediction. GBRT is a type of ensemble learning algorithm in machine learning. Compared with another ensemble learning algorithm (i.e., random forest), GBRT builds the model by adding simple regression trees into the model sequentially by boosting instead of bagging. GBRT is a strong learner that has been proved with good prediction performance (Breiman, 2001). GBRT can uncover hidden model structures in the traffic speed data to enhance the accuracy and interpretability of the model.

In the literature, traditional multi-output strategies primarily include the iterated strategy (Ikeguchi & Aihara, 1995; Williams & Zipser, 2014), direct strategy (Sorjamaa et al., 2007) and multiple-input-multiple-output (MIMO) strategy (Ben et al., 2012; Xiong et al., 2013; Bao et al., 2014; Bao et al., 2014). Both the direct strategy and iterated strategy train the samples through multi-input single-output mapping and are applicable to most prediction algorithms. For the iterated strategy, multi-step-ahead prediction requires iterating for H times, where H is the number of prediction steps. At each iteration, the predicted value of the previous model would be used as the input value of the current model to replace the corresponding input variable. Compared with the iterated strategy, the direct strategy trains H models separately and each model corresponds to an independent output. Since only observed data are used as input values, it will not cause the accumulation of errors across iterations. The MIMO strategy only trains one model to achieve multi-step-ahead prediction by considering the stochastic dependency between different prediction horizons.

Since the output values related to the same inputs are apparently autocorrelated, it's better to build a single model that is capable of simultaneously predicting traffic of all time steps. However, the tree-based ensemble methods such as GBRT cannot directly employ the

MIMO strategy for multi-step-ahead prediction due to its structure limitations. This paper proposes an MIMO strategy based on GBRT to support multi-step-ahead prediction by considering the stochastic dependency to predict all the outputs simultaneously, namely the multivariate GBRT. First, we need to store *H* output values in nodes of the regression tree which is an important part of the tree. Second, in each node, splitting criteria are applied to compute the average reduction across all *H* outputs. Through the improvement of the basic regression tree, we can extend the basic GBRT to enable the multi-step-ahead prediction.

In order to evaluate and compare the aforementioned strategies, we take into account the following three criteria: (I) The *prediction accuracy*, which determines the decision-making effectiveness of the traveler and the decision maker; (II) The *prediction stability*, which is defined as the standard deviation of the prediction errors of all prediction steps, reflects the fluctuation and discreteness of the prediction accuracy with respect to prediction steps; (III) The *prediction time*, which reflects the efficiency of the strategies. In this paper, the prediction time excludes the time for parameter-tuning and training the models, which can be conducted offline.

The contributions of this study are three folds: (I) We utilize the GBRT algorithm to achieve multi-step-ahead prediction of short-term traffic speed. Compared with one-step-ahead prediction, multi-step-ahead prediction can reflect the trend of speed change and provide more useful and meaningful information for policymakers and travelers; (II) To improve the efficiency and effectiveness, we propose the multivariate GBRT model for multi-step-ahead traffic speed forecasting; and (III) We incorporate the traditional multi-output strategies (e.g., iterated strategy, and direct strategy) into the GBRT model to achieve multi-step-ahead traffic speed forecasting. Compared to the traditional strategies, multivariate GBRT can achieve to forecast outputs of all prediction horizons simultaneously while considering the output variables' autocorrelation.

This paper is organized as follows. Section 2 revisits the literature on short-term traffic speed forecasting and reviews tree-based ensemble methods. Meanwhile, the strategies that are used to solve multi-output problems are summarized. Section 3 first presents two traditional multi-output strategies, i.e., the iterated strategy, and direct strategy, which can be used for GBRT. Then, we propose multivariate GBRT by improving the regression trees to consider the output variable autocorrelation. Section 4 compares the performances of support vector regression (SVR), direct GBRT, iterated GBRT, and multivariate GBRT in terms of the prediction accuracy, prediction time, and prediction stability, respectively. Finally, Section 5 concludes this paper and outlooks the future research.

2. LITERATURE REVIEW

In the literature, there has been extensive research conducted and numerous models proposed in the field of short-term traffic speed forecasting, e.g., historical average and smoothing methods (Farokhi Sadabadi et al., 2010; Williams et al., 1998; Guo et al., 2017), statistical and regression methods (Fei et al., 2011; Min & Wynter, 2011; Qiao et al., 2016), machine learning (Wei & Chen, 2012; Zhang & Rice, 2003), and traffic flow theory based method (Li et al., 2013; Li et al., 2014; Wu et al., 2016). Recently, artificial intelligence and machine learning algorithms have been successfully applied to the traffic prediction filed, e.g., neural networks (Hinsbergen et al., 2009; Wei & Chen, 2012), support vector machines (Wang & Shi, 2013), and hybrid or ensemble techniques (Antoniou et al., 2013). In contrast to statistical models, machine learning algorithms do not assume any specific model structures of data and treat the structure unknown. Therefore, machine learning algorithms can handle a large-size amount of data. Although the structure of the data is not apparent, machine learning algorithms can capture the potential structure. Nevertheless, a disadvantage that limits the applications of

machine learning algorithms in traffic prediction is the lack of interpretability (Vlahogianni et al., 2014).

In recent years, tree-based ensemble methods were widely used in solving prediction problems and proved to have good performance. For instance, Leshem (2007) employed the AdaBoost algorithm by taking random forests as weak base models to predict traffic flow. This algorithm could deal well with missing data. Hamner (2010) proved that random forest performed better than other models in traffic prediction for intelligent GPS navigation. Wang (2011) applied an ensemble bagging regression tree to predict the weather impact on airport capacity and demonstrated its superior performance compared with a single support vector machine. Ahmed and Abdelaty (2013) applied a stochastic gradient boosting method to study incidents and hazardous conditions, which outperformed statistical approaches. Similarly, Chung (2013) utilized boosted regression trees to identify the crash occurrence. Zhang and Haghani (2015) used the gradient boosting method in travel time prediction, but did not consider the spatial correlation between sensor locations. Considering temporal and spatial correlations, GBRT was reported to achieve a better prediction accuracy than ARIMA, generalized boosted regression models, and random forest in urban travel time forecasting (Zhang et al., 2016).

For multi-output forecasting problems, there are three commonly used strategies, i.e., iterated strategy, direct strategy, and MIMO strategy, to enable multi-step-ahead prediction. Chevillon (2007) pointed out that the iterated strategy used the predicted value as the new input, so it was easy to cause the error accumulation. In contrast to the iterated strategy that constructed a single model, the direct strategy proposed by Cox (1961) only used its past observations to build a set of models for each prediction step. So the number of models was equal to the number of prediction steps. However, there are some drawbacks on both strategies. The error accumulation in the iterated strategy greatly reduces the prediction accuracy, while the direct strategy is time-consuming and ignores the correlation between the output variables. Bontempi (2008) introduced an MIMO strategy for multi-step-ahead prediction with preserving the stochastic dependency between prediction horizons. The algorithm used for the MIMO strategy was lazy learning. Bao (2014a) proposed a PSO-MISMO modeling strategy for multi-step-ahead prediction, which was an extension of MIMO. The implementation technique for PSO-MISMO was a feed-forward neural network (FNN). As we know, the standard formulation of Support Vector Regression (SVR) can only achieve one-step-ahead prediction if not taking iterated or direct strategies. Bao (2014b) proposed the M-SVR model with the MIMO strategy for multi-step-ahead time series prediction. For the tree-based algorithm, Segal and Xiao (2011) achieved the multi-output prediction with the random forest by improving the splitting function of decision trees. Dumont et al. (2009) applied a multivariate random forest to fast multi-class image fusion, which can be used for face recognition. As an effective machine learning algorithm, there is a research need for the GBRT model with the MIMO strategy for multi-step-ahead prediction.

Multi-step-ahead prediction can provide more practical guidance to both decision makers and travelers. In this paper, we aim to achieve multi-step-ahead prediction by using GBRT with a high prediction accuracy and short forecasting time, which is not fully understood in the literature and worthy for further investigations. To the knowledge of the authors, the multi-step-ahead prediction with GBRT has not been fully studied. This paper aims to propose the multivariate GBRT that considers the stochastic dependency of outputs and predicts multi-output results simultaneously. The multivariate GBRT can achieve multi-step-ahead prediction by improving the splitting function of the basic regression tree, which will be formulated in Section 3.

3. METHODOLOGY

3.1 Gradient Boosting Regression Tree

GBRT integrates basic regression trees with the boosting method to solve the regression and classification problems. The boosting method adds additional trees in sequence, without changing the model parameters that have been added, to minimize a loss function (such as the squared error or absolute error) of the data. The loss function indicates the degree to which the prediction value deviates from the true value. By adding the basic tree that minimizes the loss function in each step, the loss function is eventually minimized.

In the regression problem, given the sample set $\{(\mathbf{x}_1, y_1),...,(\mathbf{x}_N, y_N)\}$ with N input variables $\mathbf{x}_t$ and the corresponding output variable y_t, $t = 1,..., N$. The objective is to find a function $F(\mathbf{x})$ that minimizes the specific loss function $L(y, F(\mathbf{x}))$. GBRT approaches the optimal solution $\hat{F}(\mathbf{x})$ gradually through the weighting of some simple models $h(\mathbf{x}_t)$ to minimize the loss function $L(y, F(\mathbf{x}))$. For GBRT, $h(\mathbf{x}_t)$ is the basic regression tree which is built by the input variables $\mathbf{X}$ and the negative gradient of loss function in the previous model. GBRT starts with a constant function $F_0(\mathbf{x})$ and builds the model in a greedy way. The model is formulated as follows:

$$F_0(\mathbf{x}) = \arg\min_{\gamma} \sum_{t=1}^{N} L(y_t, \gamma) \tag{1}$$

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \gamma_m h_m(\mathbf{x}) \tag{2}$$

where $F_m(\mathbf{x})$ is the integration of prediction values for the basic regression trees. $h_m(\mathbf{x})$ is the m th regression tree, γ_m is the weighting coefficient of the m th regression tree.

GBRT is optimized by the steepest descent method, the negative gradient $z_m(\mathbf{x})$ is given by

$$z_m(\mathbf{x}_t) = -\frac{\partial L(y_t, F_{m-1}(\mathbf{x}_t))}{\partial F_{m-1}(\mathbf{x}_t)} \tag{3}$$

The next regression tree $h_m(\mathbf{x})$ is built by modeling $z_m(\mathbf{x})$ and $\mathbf{X}$. The weighting coefficients can be obtained by

$$\gamma_m = \arg\min_{\gamma} \sum_{t=1}^{N} L\left(y_t, F_{m-1}(\mathbf{x}_t) - \gamma_m h_m(\mathbf{x}_t)\right) \tag{4}$$

GBRT strategically adds extra basic models to minimize the loss function. It focuses more on the samples that are difficult to estimate. In contrast to the random forest that builds each basic model by randomly sampling with an equal probability, GBRT generates the model by the boosting method. The pseudo code of GBRT is shown in Table 1.

Table 1 The pseudo code of GBRT

Input: Training set $\{(\mathbf{x}_t, y_t)\}_{t=1}^{N}$, differentiable loss function $L(y, F(\mathbf{x}))$, and the maximum number of trees M.

Initialize model with a constant value: $F_0(\mathbf{x}) = \arg\min_{\gamma} \sum_{t=1}^{N} L(y_t, \gamma)$.

For $m = 1$ to M, **do**

For $t = 1$ to N, **do**

Compute negative gradient $z_{tm} = -\left[\frac{\partial L(y_t, F_{m-1}(\mathbf{x}_t))}{\partial F_{m-1}(\mathbf{x}_t)}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}$.

End for

Fit regression tree $h_m(\mathbf{x})$ to predict negative gradient z_m using input variables $\mathbf{x}$.

Compute the gradient descent step size (learning rate) given by

$$\gamma_m = \arg\min_{\gamma} \sum_{t=1}^{N} L(y_t, F_{m-1}(\mathbf{x}_t) + \gamma h_m(\mathbf{x}_t)).$$

Update model $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \gamma_m h_m(\mathbf{x})$.

End for

Output model $F_m(\mathbf{x})$.

The performance of GBRT can be affected by the number of trees (M), learning rate (J), and max-depth of the tree (D). The best performance of the model can be achieved by selecting the best combination of these parameters via parameter tuning. M refers to the number of basic regression trees that are integrated into GBRT. With the increase in the number of trees, the error will be smaller. But too many trees will lead to overfitting and reduction in the prediction accuracy. The model will become complex and minor fluctuations in data will be exaggerated. Therefore, the number of trees needs to be controlled. J is also an important parameter, referring to the contribution of each basic regression tree to the final result. It is commonly between 0 and 1, which means to shrink the contribution of each basic regression tree. An excessive learning rate can lead to overfitting and a low learning rate may reduce the prediction accuracy. D can be expressed as the complexity of the tree. GBRT is a strong learner formed by the integration of a set of weak learners. So it is necessary to control the max-depth of each tree to limit the ability of each tree. A too large or too small D may lead to a reduction in the prediction accuracy.

3.2 Multi-Output Forecasting Strategies

3.2.1 Iterated strategy

The iterated strategy was first proposed by Chevillon (2007), which used the predicted value as the input to predict the next step output. The predicted value replaces the corresponding input value in the current iteration and the number of input variances in each iteration keeps unchanged. This step iterates for *H* times.

For the iterated strategy, given the training set $\mathcal{D} = \{(\mathbf{x}_t, y_t) \in (\mathcal{R}^d \times \mathcal{R})\}_{t=1}^{N}$, where $\mathbf{x}_t = [\varphi_t, \ldots, \varphi_{t-d+1}]^{\mathrm{T}}$ is a temporal pattern of length d, $y_t = \varphi_{t+1}$, φ_t is the time series value at time t, d represents the embedding dimension, $\mathcal{R}^d$ and $\mathcal{R}$ are the d-dimensional and 1-dimensional real sets, respectively.

The iterated strategy trains one model through the training set:

$$\varphi_{t+1} = f(\mathbf{x}_t) + \omega \tag{5}$$

where ω is the scalar zero-mean noise term.

After the training process, the predicting process is given by

$$\hat{\varphi}_{N+h} = \begin{cases} \hat{f}(\varphi_N, \varphi_{N-1}, \ldots, \varphi_{N-d+1}) & \text{if} \quad h = 1 \\ \hat{f}(\hat{\varphi}_{N+h-1}, \ldots, \hat{\varphi}_{N+1}, \varphi_N, \ldots, \varphi_{N-d+h}) & \text{if} \quad h \in \{2, \ldots, d\} \\ \hat{f}(\hat{\varphi}_{N+h-1}, \ldots, \hat{\varphi}_{N-d+h}) & \text{if} \quad h \in \{d+1, \ldots, H\} \end{cases} \tag{6}$$

As shown in Figure 1, the iterated strategy takes the predicted value as the input, which may result in the error accumulation. When the number of prediction steps increases, the effect of the error accumulation will be enlarged. So this strategy performs poor in terms of the multi-step-ahead prediction accuracy. This strategy only needs to establish one model to predict but it cannot output all the results simultaneously.

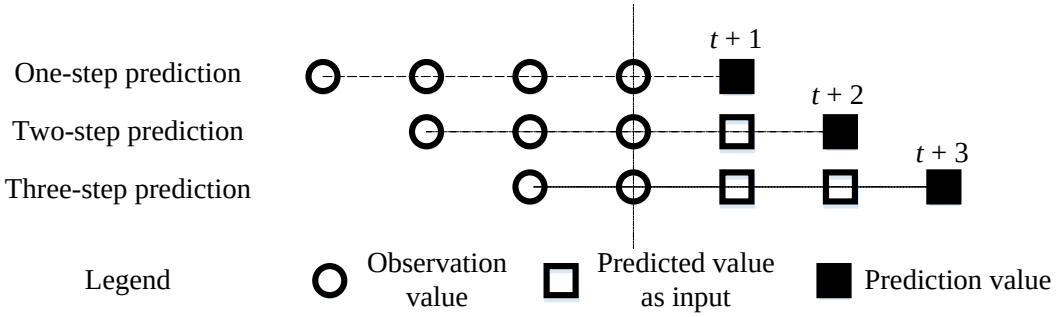


Figure 1 An illustration of the iterated strategy.

3.3.2 Direct strategy

The direct strategy was first proposed by Cox (1961), which required building a set of models for each step. The input variable only takes into account the actually observed values rather than the predicted values. So the model for each step is relatively independent.

For the direct strategy, given the training set: $\mathcal{D}_1 = \{(\mathbf{x}_t, y_{t1}) \in (\mathcal{R}^d \times \mathcal{R})\}_{t=1}^N$,…, $\mathcal{D}_H = \{(\mathbf{x}_t, y_{tH}) \in (\mathcal{R}^d \times \mathcal{R})\}_{t=1}^N$, where $\mathrm{x}_t \in \{(\varphi_t, \ldots, \varphi_{t-d+1})\}$, $\mathrm{y}_{th} = \varphi_{t+h}$, for which we intend to predict the next H observations by using the single-output method.

The direct strategy trains H models through the training sets $\mathcal{D}_h \in \{\mathcal{D}_1, \ldots, \mathcal{D}_H\}$:

$$\varphi_{t+h} = f_h(\mathbf{x}_t) + \omega \tag{7}$$

where f_h is the learned single-output model, $h = 1, \ldots, H$.

After the training process, the predicting process is as follows:

$$\hat{\varphi}_{N+h} = \hat{f}_h(\varphi_N, \varphi_{N-1}, \ldots, \varphi_{N-d+1}), \qquad h \in \{1, \ldots, H\} \tag{8}$$

As shown in Figure 2, in contrast to the iterated strategy, the direct strategy is not easy to form the error accumulation. However, this strategy ignores the correlation between output variables. Since it needs to train a set of models independently for each step of the outputs, more computational costs will be spent on the training process.

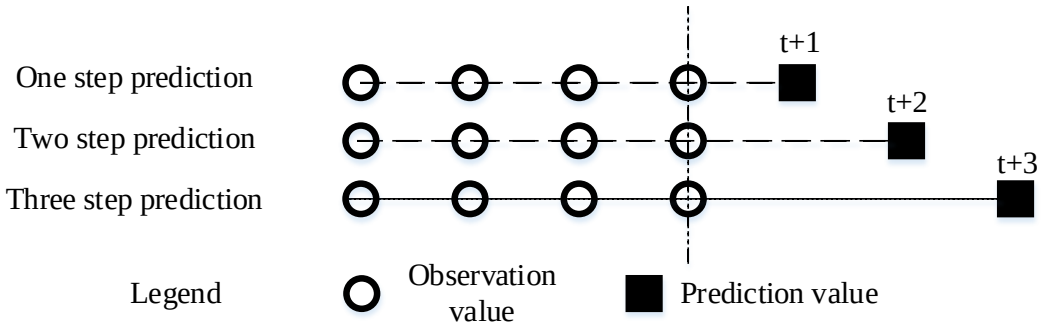


Figure 2 An illustration of the direct strategy.

3.3 Multivariate GBRT

The multivariate GBRT can consider the correlation of the output variables and predict all output variables simultaneously. The basic GBRT model adds basic decision trees in sequence. In order to output all prediction values simultaneously, we need to improve the basic GBRT. The main idea is to store all the output variables simultaneously rather than one output in the nodes of the regression tree. Besides, in the process of splitting the nodes into child nodes, the splitting function takes into account all the output variables.

The regression tree has the tree-based structure whose basic elements include the internal nodes, branch and leaf nodes. Each internal node needs to perform a binary split (yes/no). Each branch represents a splitting output and each leaf node represents a splitting category. The most important part of the regression tree is the splitting process. The splitting process makes the nodes to generate child nodes. At the same time, the samples are passed from nodes to the child nodes. With the splitting process, the regression trees are gradually developed and eventually reach the leaf nodes (terminal nodes). The first step in the splitting process is to establish a series of binary questions (yes/no) based on the input features. Those questions determine how the sample is allocated to the child nodes and divide the sample into two parts. The second step is to use the impurity measure as the criterion of splitting for each node. Generally, the impurity measure in the regression problem is the variance of the output variables. Finally, the splitting function $\phi(s,k)$ is used to evaluate the results of each allowable splitting s at the node k. The optimal splitting in all feasible splits corresponds to the optimal splitting function, which means that the sample distribution in the child nodes for the optimal splitting is the most uniform. The procedure of splitting process is shown in Figure 3.

For single-output problems, given the sample set $\{(\mathbf{x}_1, y_1),...,(\mathbf{x}_N, y_N)\}$ with input variables $\mathbf{x}_t$ and the corresponding prediction variable y_t, $t = 1,..., N$. The splitting process divides the nodes into two child nodes, namely the left node k_L and a right node k_R. We try all the possible splits of the node k based on the binary splitting question. The impurity measure of the node k is $SS(k) = \sum_{t\in k}(y_t - \mu(k))^2$, where $\mu(k)$ is the mean of y_t for all the samples in the node k. The corresponding splitting function is $\phi(s,k) = SS(k) - SS_L(k) - SS_R(k)$. Through the above splitting process, the regression tree can achieve the single output.

For multi-output problems, given the sample set $\{(\mathbf{x}_1, \mathbf{y}_1),...,(\mathbf{x}_N, \mathbf{y}_N)\}$ with input variables $\mathbf{x}_t$ and the corresponding prediction variables $\mathbf{y}_t = [y_{t1}, y_{t2}, \ldots, y_{tH}]^{\mathrm{T}}$, $t = 1,..., N$. The prediction variable is an $H \times 1$ vector. Binary splitting questions are based on the input

features, where the multivariate GBDT needs to be improved is impurity measure. First, we need to define the correlation matrix of the prediction variables $\mathbf{V}(\theta,k)$ in node k, where θ is the prediction variable sequence length when calculating the correlation coefficient. That is to calculate the correlation coefficients of the sequences for each prediction step. When θ equals to $H(H+1)/2$, we can make no assumptions on the correlation matrix. In order to consider the interpretability and efficiency of the model, we need to limit the value of θ, generally taking the smaller θ (Segal, 1992).

In order to achieve multiple outputs, the impurity measure can be improved as follows:

$$SS(k)=\sum_{t=1}^{N}(\mathbf{y}_t-\boldsymbol{\mu}(k))^{\mathrm{T}}\mathbf{V}^{-1}(\theta,k)(\mathbf{y}_t-\boldsymbol{\mu}(k)) \tag{9}$$

where $\mathbf{V}(\theta,k)$ represents the correlation matrix for each prediction step, θ is the sequence length when calculating the correlation coefficient, k is the node number, and $\boldsymbol{\mu}(k)$ is the $H\times 1$ mean vector of the predictor variables in the node k.

The form of the multi-output splitting function is the same as the single-output splitting function, namely $\phi(s,k)=SS(k)-SS_L(k)-SS_R(k)$. Through the above improvement, the regression tree can output all prediction variances simultaneously. Thus, the multivariate GBDT is used to achieve multi-step-ahead prediction simultaneously by the boosting method to integrate basic decision trees. Since the correlation matrix is introduced to calculate the impurity measure, the correlation between the output variables is also considered during the calculation.

As shown in Table 2, the three strategies have different characteristics in training and predicting process and can be used to solve multi-output prediction problems.

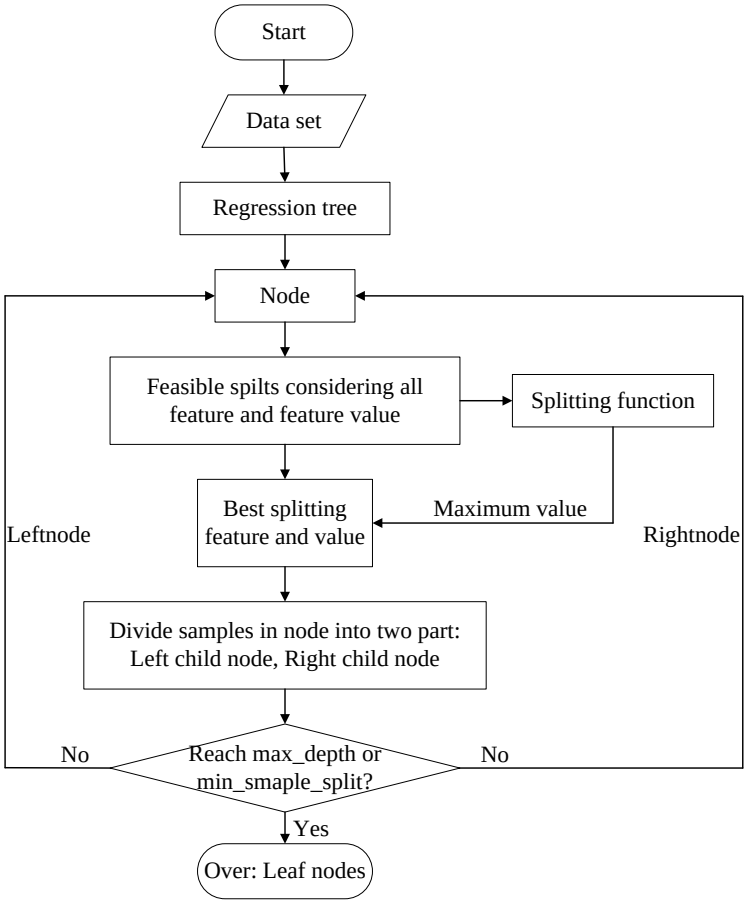


FIGURE 3 A flowchart of the splitting process for basic regression trees.

Table 2 Comparison of the three multi-output strategies

Multi-output GBRT	Model number	Training times	Predicting times	Whether using the predicted value as inputs?	Remark
Direct strategy	*H*	*H*	*H*	yes	Ignoring correlations among outputs
Iterated strategy	1	1	*H*	no	Error accumulation
Multivariate GBRT	1	1	1	yes	Predict all outputs simultaneously

Note: *H* is the prediction horizons.

4. EXPERIMENTS

In this paper, SVR is used as the benchmark model because it is widely used in the field of prediction due to a good prediction accuracy and GBRT is taken as the basic model. Several multi-output traffic prediction models are tested and compared to achieve multi-step-ahead prediction, i.e., GBRT using the direct strategy, GBRT using the iterated strategy, and the proposed multivariate GBRT. The performance of those three strategies is evaluated from the three perspectives: prediction accuracy, prediction stability, and prediction time.

In order to test the performance of the three strategies at different prediction horizons, two experiments are conducted in this paper, i.e., 6-step-ahead prediction (30-min), and 12-step-ahead prediction (60 min).

4.1 Data Description and Preparation

In this paper, traffic data are extracted at the 5-min aggregation time interval from three loop detectors in US101-N freeway between May 1 and June 14, 2017 (45 days), through the Performance Measurement System (PeMS) that provides real-time and historical data collected from over 42,000 detectors deployed on freeways throughout California. The observation rate of the extracted data at the study site is up to 99.6%, which means only 0.4% of data are missed and repaired by imputation method (temporal medians).

The three loop detectors are located from the 3rd Street to Bacon Street in the US101-N. The first 31-day data are used as the training set, the middle 7-day data as the validation set, and the remaining 7-day data are used as the test set. The 5-min aggregated travel speed is used to represent traffic conditions. As shown in Figure 4, the detection locations are denoted by W2 (upstream), W3 (target), and W4 (downstream), respectively. The distance between W2 and W3 is 0.45 miles and the distance between W3 and W4 is 0.3 miles. To consider the spatial correlation, the speed data from the three loop detectors are used. That is, we aim to predict the speed of W3 by incorporating data collected from both W2 and W4.

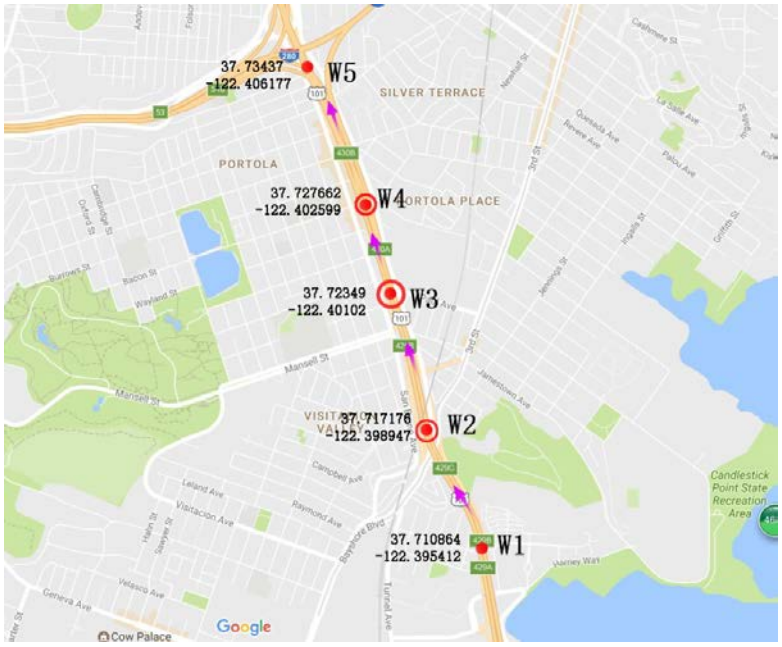


FIGURE 4 Configuration of loop detectors on US101-N, CA.

Figure 5 shows that the patterns of traffic speed at target detector W3 are distinct in different days. There are clear morning and evening peaks on Wednesday while the peak hours happen in the afternoon for Saturday and Sunday. The pattern of speed change on Monday is a little different from that of Wednesday, on which the evening peak is not obvious.

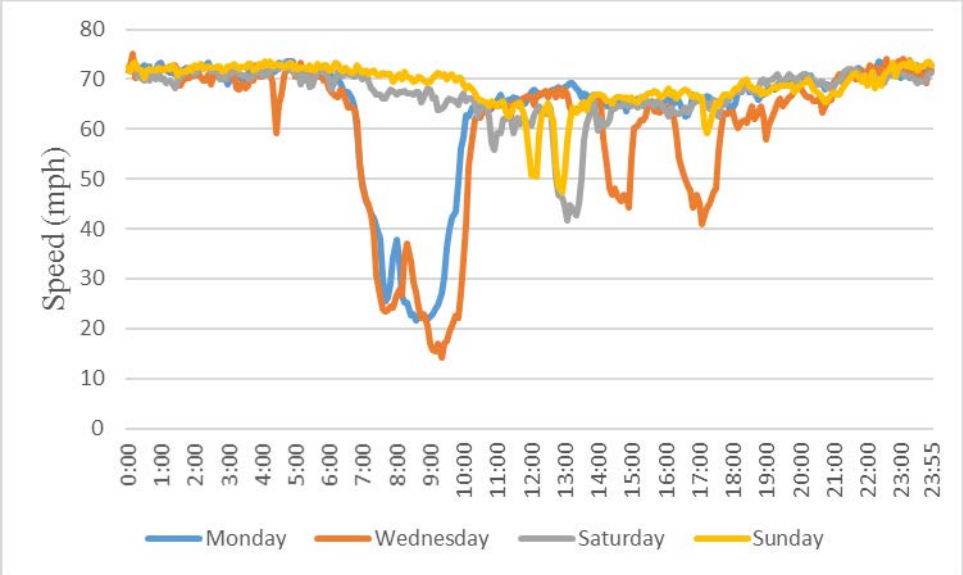


FIGURE 5 Temporal traffic speed profiles at target detector W3.

Table 3 proves that the speed at W3 has a great correlation with the speed at its upstream and downstream. When the coefficient of correlation is close to 1, the correlation between two detectors is close to perfectly positive correlation. As shown in Table 4, the speeds of detectors have significant temporal correlations with their previous traffic speeds. So considering the spatial and temporal correlation can improve the speed prediction accuracy.

Table 3 Coefficient of correlation of speed among target detector W3, upstream detector W2, and downstream detector W4 for a week

Detector	Correlation coefficient
W2-W3	0.843**
W3-W4	0.933**

** Significantly correlated at the 0.01 confidence level (two-tailed). * Significantly correlated at the 0.05 confidence level (two-tailed).

Table 4 Temporal autocorrelation of target detector W3 for different lag times

Detector	t	t-1	t-2	t-3	t-4	t-5
W2	1	0.974**	0.925**	0.864**	0.801**	0.739**
W3	1	0.987**	0.965**	0.938**	0.910**	0.883**
W4	1	0.984**	0.960**	0.933**	0.907**	0.882**

** Significantly correlated at the 0.01 confidence level (two-tailed). * Significantly correlated at the 0.05 confidence level (two-tailed).

Table 5 lists basic characteristics of the inputs variables/features for multi-step-ahead traffic speed forecasting. In order to consider both the spatial and temporal correlations, the speed data at the upstream and downstream locations are used as input variables and the speed within the previous 25-min time interval is taken into account. Since the data in the training set, validation set, and test set are all larger than 7 days, traffic speed differences between weekends and weekdays have been considered.

Table 5 Features for multi-step-ahead traffic speed forecasting

ID	Features	Feature Description	Type	Range/set
1	Day of week	Day of week	discrete	{1, 2, 3,…, 7}*
2	Time of day	Every 5-min time stamp of the day	discrete	{1, 2, 3,…, 288}**
3	j_{t-1}	Speed measurements of the upstream detector at time step $t-1$	continuous	[0, +∞)
4	j_{t-2}	Speed measurements of the upstream detector at time step $t-2$	continuous	[0, +∞)
5	j_{t-3}	Speed measurements of the upstream detector at time step $t-3$	continuous	[0, +∞)
6	j_{t-4}	Speed measurements of the upstream detector at time step $t-4$	continuous	[0, +∞)
7	j_{t-5}	Speed measurements of the upstream detector at time step $t-5$	continuous	[0, +∞)
8	f_{t-1}	Speed measurements of the target detector at time step $t-1$	continuous	[0, +∞)
9	f_{t-2}	Speed measurements of the target detector at time step $t-2$	continuous	[0, +∞)
10	f_{t-3}	Speed measurements of the target detector at time step $t-3$	continuous	[0, +∞)
11	f_{t-4}	Speed measurements of the target detector at time step $t-4$	continuous	[0, +∞)

12	f_{t-5}	Speed measurements of the target detector at time step $t-5$	continuous	[0, +∞)
13	h_{t-1}	Speed measurements of the downstream detector at time step $t-1$	continuous	[0, +∞)
14	h_{t-2}	Speed measurements of the downstream detector at time step $t-2$	continuous	[0, +∞)
15	h_{t-3}	Speed measurements of the downstream detector at time step $t-3$	continuous	[0, +∞)
16	h_{t-4}	Speed measurements of the downstream detector at time step $t-4$	continuous	[0, +∞)
17	h_{t-5}	Speed measurements of the downstream detector at time step $t-5$	continuous	[0, +∞)
18	Δ_{t-1}	$\Delta_{t-1} = f_{t-1} - f_{t-2}$	continuous	(-∞, +∞)
19	Δ_{t-2}	$\Delta_{t-2} = f_{t-2} - f_{t-3}$	continuous	(-∞, +∞)
20	Δ_{t-3}	$\Delta_{t-3} = f_{t-3} - f_{t-4}$	continuous	(-∞, +∞)
21	Δ_{t-4}	$\Delta_{t-4} = f_{t-4} - f_{t-5}$	continuous	(-∞, +∞)
22	f_t	Observed travel speed at time step t	continuous	[0, +∞)

Note: * Numbers 1,…, 7 indicate Monday through Sunday; * Numbers 1,…, 288 indicate each 5 min.

Figure 6 lists the importance rankings of all features used in the model and the parameter combination is adopted as described in Section 4.3. We can find that f_{t-1} has the greatest impact on the final prediction results and the feature of time of day also has the high relative importance. The features with the spatial correlations like j_{t-1} and h_{t-1} have non-negligible impacts on the model. The differences in speed of two adjacent time stamps like Δ_{t-1} are important for the model. As shown in Figure 6, all features used in this model have considerable contributions to the final outputs.

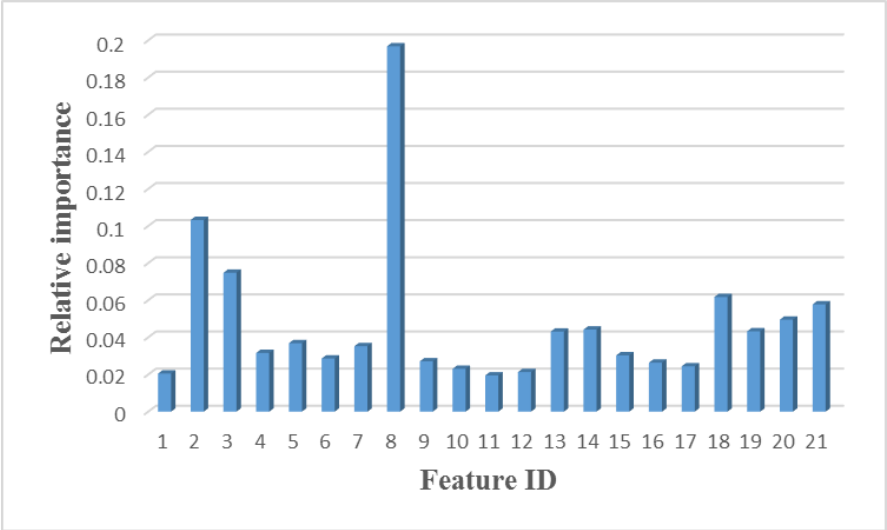


FIGURE 6 Relative importance ranking of features used in the model.

4.2 Measures of Effectiveness

In this paper, the prediction accuracy, prediction time, and prediction stability are used as the evaluation criteria. Through a comprehensive comparison, the performance of the three strategies in multi-step-ahead prediction can be clearly illustrated.

4.2.1 Prediction accuracy

In order to compare the advantages and disadvantages of different strategies, four error measures are defined as follows:

(I) Mean absolute percentage error (MAPE)

$$\mathrm{MAPE} = \frac{1}{m}\sum_{t=1}^{m}\left|\frac{f_t - \hat{f}_t}{f_t}\right| \times 100\% \tag{10}$$

(II) Symmetric mean absolute percentage error (SMAPE)

$$\mathrm{SMAPE1} = \frac{1}{m}\sum_{t=1}^{m}\frac{\left|f_t - \hat{f}_t\right|}{(f_t + \hat{f}_t)/2} \times 100\% \tag{11}$$

$$\mathrm{SMAPE2} = \frac{\sum_{t=1}^{m}\left|f_t - \hat{f}_t\right|}{\sum_{t=1}^{m}(f_t + \hat{f}_t)/2} \times 100\% \tag{12}$$

(III) Root mean square error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{t=1}^{m}(f_t - \hat{f}_t)^2} \tag{13}$$

(IV) Normalized root mean square error (NRMSE)

$$\mathrm{NRMSE} = \sqrt{\frac{\sum_{t=1}^{m}(f_t - \hat{f}_t)^2}{\sum_{t=1}^{m} f_t^2}} \times 100\% \tag{14}$$

where f_t denotes the true speed value at time t, $\hat{f}_t$ denotes the predicted speed value at time t, and m is the sample size of the test set.

4.2.2 Prediction stability

For multi-step-ahead prediction, this paper introduces a new evaluation criterion, namely, prediction stability, which indicates the standard deviation of prediction errors for multiple steps. The stability of different strategies in multi-step-ahead prediction is determined by calculating the standard deviation of the prediction accuracy of all steps. The smaller the standard deviation, the higher the prediction stability. That is, the transition between the synchronization accuracy is more stable. With the increase of prediction horizons, the model with higher stability will perform better.

4.2.3 Prediction time

In field applications, decision makers and travelers pay more attention to the real-time

prediction time because parameter tuning and model training can be conducted periodically offline. Thus, parameter tuning and training time have no direct impacts on real-time prediction time. So this paper only considers the real-time prediction time. Too long prediction time will seriously affect the user experience. While a short prediction time is helpful to accelerate the prediction information provision, improve the user experience, increase the applicability of the prediction model, and reduce delay.

4.3 Parameter Tuning

In this paper, the grid search method is used to search for the optimal parameters. The evaluation criterion for parameter tuning is minimizing MAPE. SVR uses the radial basis function (RBF) as the kernel function. So the parameters that need to be adjusted are C and γ, where C is the penalty factor that represents the tolerance to the error. A higher C value means a lower tolerance to the error. γ is a parameter that comes with the RBF, which implicitly determines the distribution of the data in the new feature space. Through the grid search, 36 combinations of parameters C and γ are evaluated with each ranging from 10^{-3} to 10^{3}. The optimal combination of the parameters obtained from the training set and validation set would be used to predict speed in the test set.

As depicted in Section 3.1, we evaluate 448 combinations of GBRT parameters by the grid search method, too. The GBRT model is trained with M ranging from 1 to 5000, J from 0.0005 to 0.5 and D from 1 to 8. The step sizes of grid search for GBRT are shown in Figure 7. For GBRT, there are three strategies to achieve multi-step-ahead prediction including the direct GBRT, iterated GBRT, and multivariate GBRT. Considering the similarity of the data structure and parameter tuning time, parameter tuning is only conducted for the direct GBRT and iterated GBRT. The parameter tuning results of the first-step prediction model are employed in the following multi-ahead-prediction models. Multi-output GBRT can output all prediction values simultaneously so only one shot of parameter tuning is required.

For the 30-min-ahead prediction experiment, Figure 7 shows the relationship between the prediction accuracy of models and the number of trees with different learning rates (J) and maximum depth (D) for multivariate GBRT. Before a certain threshold (number of trees), MAPE decreases with M, while MAPE increases when M exceeds the threshold. The decreasing slopes of the curves vary with the different learning rates. So the prediction accuracy will be the best when M reaches the threshold. As shown in Figure 8, a certain value of D will reach the optimum, not necessarily the highest maximum depth. Table 6 shows the optimal parameter combination for different strategies and the corresponding prediction accuracy. The MAPE of multivariate GBRT represents the total error of 6-step predictions while the MAPE of SVR and other GBRT strategies is the error of the first-step prediction only. So the MAPE of multivariate GBRT is relatively high.

For the 60-min-ahead prediction experiment, Table 7 shows the optimal combination of parameters and the corresponding prediction accuracy. In this experiment, the optimal combination of parameters for multivariate GBRT and other strategies are different. From the optimal combination of parameters, we can know that the optimal maximum depth increases when the sample structure becomes more complex. For example, the optimal maximum depth increase from 5 to 7 when the prediction step increases from 6 to 12.

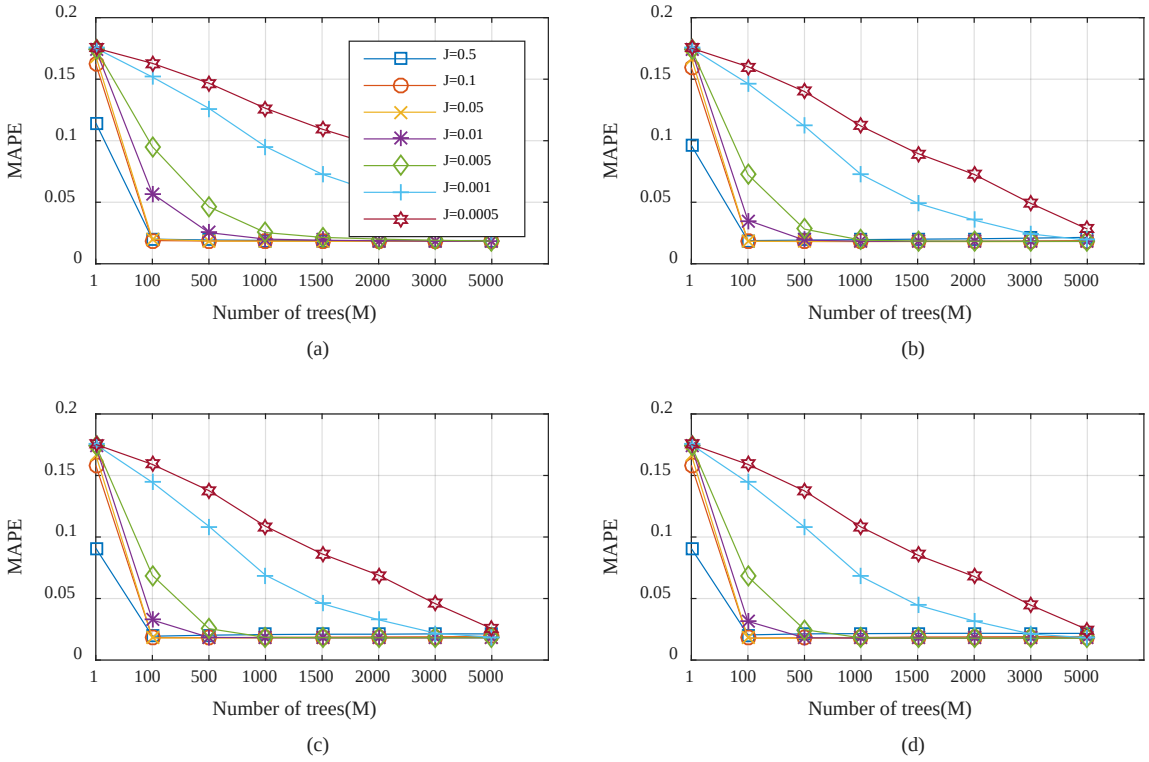


Figure 7 The relationship between MAPE and the number of trees (*M*) with different learning rates (*J*) and maximum depths (*D*): (a) *D* = 1; (b) *D* = 2; (c) *D* = 3; (d) *D* = 4.

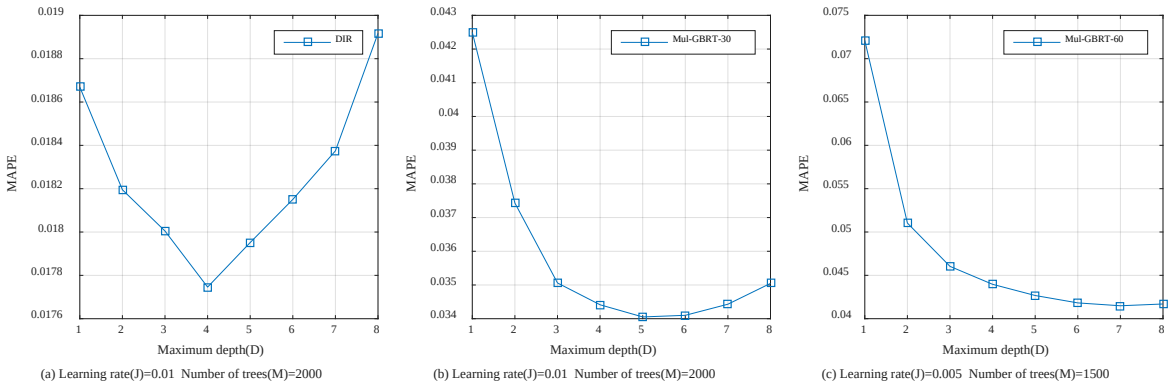


Figure 8 The relationship between MAPE and the maximum depth (*D*) with different learning rates (*J*) and numbers of trees (*M*).

Table 6 Optimal combination of parameters for 30-min-ahead prediction

Model	*M*	*J*	*D*	MAPE
Multivariate GBRT	2000	0.01	5	3.41%
Direct GBRT	2000	0.01	4	**1.77%**
Iterated GBRT	2000	0.01	4	**1.77%**
-	*C*	γ	-	MAPE
SVR	10	0.001	-	2.02%

Table 7 Optimal combination of parameters for 60-min-ahead prediction

Model	*M*	*J*	*D*	MAPE
Multivariate GBRT	1500	0.005	7	4.15%
Direct GBRT	2000	0.01	4	1.77%
Iterated GBRT	2000	0.01	4	1.77%
-	C	γ	-	MAPE
SVR	10	0.001	-	2.02%

As shown in Table 8, the parameter tuning time is different for the models. The total tuning time is the whole computing time for all parameter combinations while the average tuning time is equal to the total tuning time divided by the number of combinations. We can find that the average tuning time is quite short for each model and the values are similar. To accelerate the procedure of parameter tuning, parallel computing and narrowing grid search space can be adopted in the real-world operation.

Table 8 Comparison of parameter tuning time

Model	Total tuning time (h)	Average tuning time (s)
SVR	0.40	29.06
Direct GBRT	3.39	27.21
Iterated GBRT	3.39	27.21
Multivariate GBRT (30-min-ahead)	4.14	33.25
Multivariate GBRT (60-min-ahead)	6.18	49.67

4.4 Model Comparison

In this section, the aforementioned models are evaluated and compared with the measures of effectiveness defined in Section 4.2. All models in this section are executed 10 times and the results is obtained from the average values of 10 execution results. To ensure the statistical significance of the models comparison, analysis of variance (ANOVA) test and Tukey's test are employed. ANOVA is used to determine if the results of four models are statistically different. We use Tukey's test to compare the results of all pairwise models and Tukey's test is a post-hoc test which means it can be perform only if we reject the null hypothesis in ANOVA.

4.4.1 Prediction accuracy

We employ the ANOVA test on the results of prediction accuracy and the analysis results indicate that there are statistically differences among four models at 0.05 significance level. Then by using Tukey's test, we check the statistically differences between each pairwise models. The analysis results prove the error differences between any two models are significant at 0.05 significance level, which mean the ranks in Table 9 and Table 10 are credible. The analysis results are not listed alone limited by the length of paper.

Table 9 compares the prediction accuracies of the models. All of the direct GBRT, iterated GBRT and multivariate GBRT are better than SVR in 30-min-ahead prediction. Iterated GBRT has the best performance in accuracy because it uses updated predicted values as inputs. It strengthens the relationship between adjacent steps while the effect of error accumulation is not significant in short-step prediction. In 30-min-ahead prediction, the direct GBRT has the

better prediction performance than multivariate GBRT but the margin of the prediction accuracy is quite small. Both of them have good prediction accuracies.

As shown in Table 10, the prediction accuracy of iterated GBRT decreases rapidly in 60-min-ahead prediction, which is very sensitive to the error accumulation in the iteration process. The prediction accuracy of direct GBRT and multivariate GBRT are significantly better than the performance of SVR, in which multivariate GBRT has a large degree of improvement in prediction accuracy compared with the 30-min-ahead prediction. The prediction accuracies of direct GBRT and multivariate GBRT are comparable. In general, both multivariate GBRT and direct GBRT have a better prediction accuracy than SVR in 60-min-ahead prediction. The advantage of multivariate GBRT in multi-step-ahead prediction is gradually reflected with the increasing prediction step. Iterated GBRT is greatly affected by the error accumulation which leads to a poor accuracy.

Table 9 Comparison of prediction accuracy for 30-min-ahead prediction

Model	Error	Rank	Prediction step					
			1	2	3	4	5	6
	MAPE							
SVR	5.12%	4	2.75%	3.97%	5.09%	5.83%	6.35%	6.75%
Direct GBRT	4.64%	2	2.06%	3.42%	4.52%	5.36%	5.99%	6.48%
Iterated GBRT	4.50%	1	2.06%	3.24%	4.32%	5.17%	5.85%	6.34%
Multivariate GBRT	4.70%	3	2.22%	3.48%	4.59%	5.43%	6.03%	6.49%
	SMAPE1							
SVR	4.70%	4	2.65%	3.78%	4.72%	5.32%	5.73%	6.01%
Direct GBRT	4.38%	2	2.04%	3.32%	4.30%	5.02%	5.59%	6.01%
Iterated GBRT	4.23%	1	2.04%	3.16%	4.13%	4.86%	5.42%	5.79%
Multivariate GBRT	4.45%	3	2.19%	3.37%	4.35%	5.10%	5.63%	6.03%
	SMAPE2							
SVR	3.74%	4	2.12%	3.01%	3.75%	4.21%	4.55%	4.80%
Direct GBRT	3.62%	2	1.75%	2.76%	3.55%	4.14%	4.57%	4.92%
Iterated GBRT	3.48%	1	1.75%	2.62%	3.38%	3.95%	4.40%	4.73%
Multivariate GBRT	3.66%	3	1.84%	2.78%	3.60%	4.20%	4.62%	4.93%
	RMSE							
SVR	4.36	4	2.27	3.29	4.18	4.78	5.19	5.55
Direct GBRT	4.09	1	1.75	3.02	3.90	4.50	4.96	5.30
Iterated GBRT	4.32	3	1.75	2.94	3.94	4.70	5.35	5.85
Multivariate GBRT	4.20	2	1.86	3.09	4.04	4.65	5.07	5.41
	NRMSE							
SVR	6.85%	4	3.56%	5.17%	6.56%	7.52%	8.16%	8.72%
Direct GBRT	6.43%	1	2.75%	4.74%	6.14%	7.07%	7.80%	8.33%
Iterated GBRT	6.79%	3	2.75%	4.63%	6.19%	7.39%	8.40%	9.19%
Multivariate GBRT	6.60%	2	2.92%	4.86%	6.35%	7.30%	7.96%	8.49%

Table 10 Comparison of prediction accuracy for 60-min-ahead prediction

Model	Error	Rank	Prediction step											
			1	2	3	4	5	6	7	8	9	10	11	12
	MAPE													
SVR	6.34%	4	2.75%	3.97%	5.09%	5.83%	6.35%	6.75%	6.93%	7.09%	7.37%	7.76%	8.04%	8.22%
Direct GBRT	5.94%	2	2.06%	3.42%	4.52%	5.36%	5.99%	6.48%	6.81%	7.09%	7.32%	7.34%	7.41%	7.45%
Iterated GBRT	6.01%	3	2.06%	3.24%	4.32%	5.17%	5.85%	6.34%	6.78%	7.12%	7.38%	7.65%	7.94%	8.25%
Multivariate GBRT	5.92%	1	2.44%	3.67%	4.64%	5.37%	5.94%	6.41%	6.70%	6.89%	7.05%	7.15%	7.29%	7.45%
	SMAPE1													
SVR	5.71%	4	2.65%	3.78%	4.72%	5.32%	5.73%	6.01%	6.18%	6.32%	6.58%	6.88%	7.09%	7.20%
Direct GBRT	5.55%	2	2.04%	3.32%	4.31%	5.02%	5.59%	6.01%	6.31%	6.56%	6.76%	6.81%	6.90%	6.93%
Iterated GBRT	5.53%	1	2.04%	3.16%	4.13%	4.86%	5.41%	5.79%	6.16%	6.47%	6.71%	6.96%	7.20%	7.42%
Multivariate GBRT	5.57%	3	2.42%	3.58%	4.45%	5.09%	5.59%	6.00%	6.26%	6.45%	6.60%	6.69%	6.80%	6.93%
	SMAPE2													
SVR	4.58%	4	2.12%	3.01%	3.75%	4.21%	4.55%	4.80%	4.95%	5.19%	5.30%	5.52%	5.69%	5.79%
Direct GBRT	4.57%	3	1.75%	2.76%	3.55%	4.14%	4.57%	4.91%	5.17%	5.43%	5.57%	5.61%	5.68%	5.70%
Iterated GBRT	4.54%	1	1.75%	2.62%	3.38%	3.95%	4.40%	4.73%	5.06%	5.33%	5.53%	5.73%	5.90%	6.07%
Multivariate GBRT	4.55%	2	1.98%	2.92%	3.64%	4.14%	4.55%	4.89%	5.11%	5.27%	5.39%	5.47%	5.54%	5.64%
	RMSE													
SVR	5.30	3	2.27	3.29	4.18	4.78	5.19	5.55	5.67	5.81	6.00	6.22	6.36	6.47
Direct GBRT	5.09	1	1.75	3.02	3.90	4.50	4.95	5.30	5.55	5.85	5.98	6.07	6.04	6.07
Iterated GBRT	5.63	4	1.75	2.94	3.93	4.70	5.33	5.83	6.20	6.48	6.67	6.79	6.90	7.05
Multivariate GBRT	5.15	2	2.08	3.30	4.07	4.55	4.95	5.34	5.66	5.86	5.97	6.01	6.05	6.12
	NRMSE													
SVR	8.32%	3	3.56%	5.17%	6.56%	7.52%	8.16%	8.72%	8.91%	9.12%	9.42%	9.77%	10.00%	10.17%
Direct GBRT	8.00%	1	2.75%	4.74%	6.13%	7.07%	7.78%	8.33%	8.72%	9.19%	9.39%	9.53%	9.49%	9.54%
Iterated GBRT	8.84%	4	2.75%	4.62%	6.18%	7.38%	8.38%	9.16%	9.75%	10.18%	10.47%	10.66%	10.77%	10.99%
Multivariate GBRT	8.08%	2	3.26%	5.18%	6.40%	7.15%	7.78%	8.39%	8.89%	9.21%	9.38%	9.44%	9.51%	9.62%

4.4.2 Prediction stability

ANOVA test and Tukey's test are also adopted on the results of prediction stability. The analysis results indicate that there are statistically significant differences among four models in prediction stability and the ranks of prediction stability is believable through pairwise model comparisons.

The prediction stability analysis is to determine the smoothness of the prediction variation in the multi-step-ahead prediction by calculating the standard deviation of each step prediction error. A higher stability can prevent the prediction accuracy from suddenly decreasing in the relatively long-term prediction.

Table 11 Comparison of prediction stability

Model	30-min-ahead prediction		60-min-ahead prediction	
	Prediction stability	Rank	Prediction stability	Rank
	MAPE		MAPE	
SVR	1.52%	1	1.68%	2
Direct GBRT	1.67%	4	1.77%	3
Iterated GBRT	1.63%	3	1.96%	4
Multivariate GBRT	1.62%	2	1.59%	1
	SMAPE1		SMAPE1	
SVR	1.28%	1	1.39%	1
Direct GBRT	1.49%	4	1.59%	3
Iterated GBRT	1.43%	3	1.69%	4
Multivariate GBRT	1.46%	2	1.43%	2
	SMAPE2		SMAPE2	
SVR	1.02%	1	1.13%	1
Direct GBRT	1.19%	4	1.29%	3
Iterated GBRT	1.14%	2	1.37%	4
Multivariate GBRT	1.18%	3	1.17%	2
	RMSE		RMSE	
SVR	1.24	1	1.31	2
Direct GBRT	1.33	3	1.39	3
Iterated GBRT	1.55	4	1.72	4
Multivariate GBRT	1.33	2	1.28	1
	NRMSE		NRMSE	
SVR	1.95%	1	2.05%	2
Direct GBRT	2.09%	2	2.19%	3
Iterated GBRT	2.42%	4	2.67%	4
Multivariate GBRT	2.10%	3	2.02%	1

Table 11 shows that SVR has the strongest prediction stability in both 30-min-ahead and 60-min ahead prediction, which indicates that its prediction accuracy is the most stable with multiple prediction steps. Among GBRT models, multivariate GBRT has the highest prediction stability. In 60-min ahead prediction, iterated GBRT has the worst prediction stability because of the rapid accuracy descent caused by the error accumulation. Multivariate

GBRT outputs all the prediction values simultaneously. Although the prediction accuracy of multivariate GBRT is a little bit worse than direct GBRT and iterated GBRT in the first-step prediction, with the increasing prediction step, the gap in the prediction accuracy among these multi-output GBRT models gradually becomes narrowed. It can be seen that the advantages of the prediction stability can be reflected in the prediction with long steps. Besides, the prediction stability is also an important consideration for travelers and traffic managers.

4.4.3 Prediction time

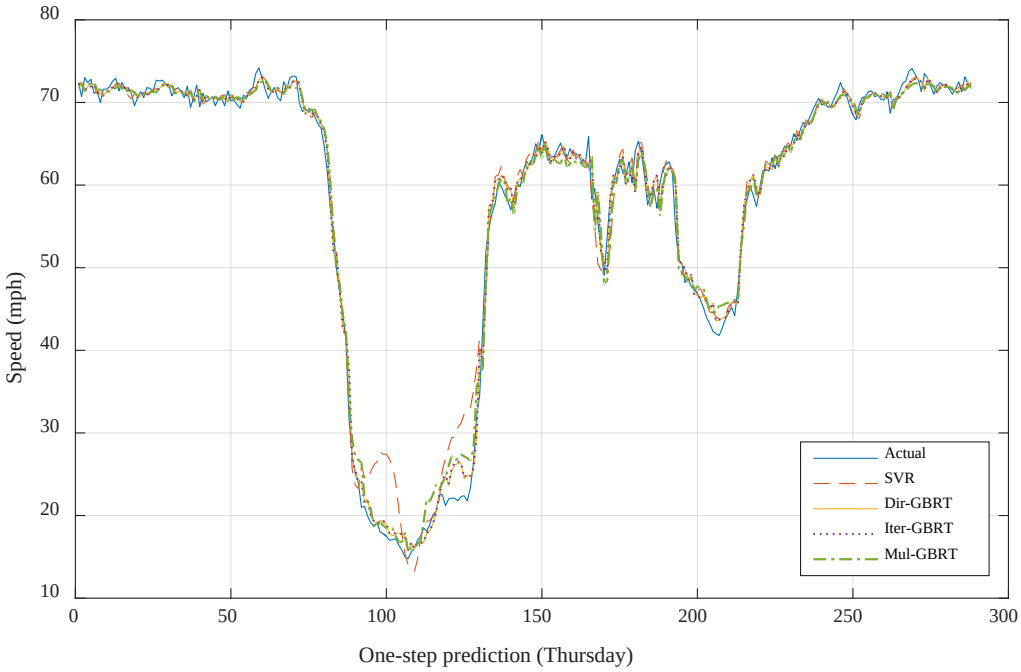
The prediction time only considers the predicting process, regardless of the parameter tuning and training process. In the real-time applications, the prediction time determines the prediction efficiency. A shorter prediction time means a better applicable potential of the model and better user experience. As shown in Table 12, multivariate GBRT has significant advantages in the prediction time in 30-min-ahead prediction and the prediction time is far less than other models. Since the direct GBRT and iterated GBRT need to predict each step separately in the prediction process, the prediction time is greater than multivariate GBRT. Iterated GBRT needs to update the feature space in each iteration process and replaces the corresponding values with predicted values. Due to the spatial and temporal correlations, it is necessary to predict the speeds of the upstream and downstream sections in each iteration and replace them with predicted values. So the prediction time of iterated GBRT is 3 times longer than direct GBRT. SVR has no advantage in prediction time because the prediction time of SVR is much longer than that of GBRT. For 60-min-ahead prediction, with the increase of the prediction step, the gap in prediction time between multivariate GBRT and the other GBRT models is gradually widened. SVR is time-consuming so it may have limitations in real-time applications.

Table 12 Comparison of prediction time

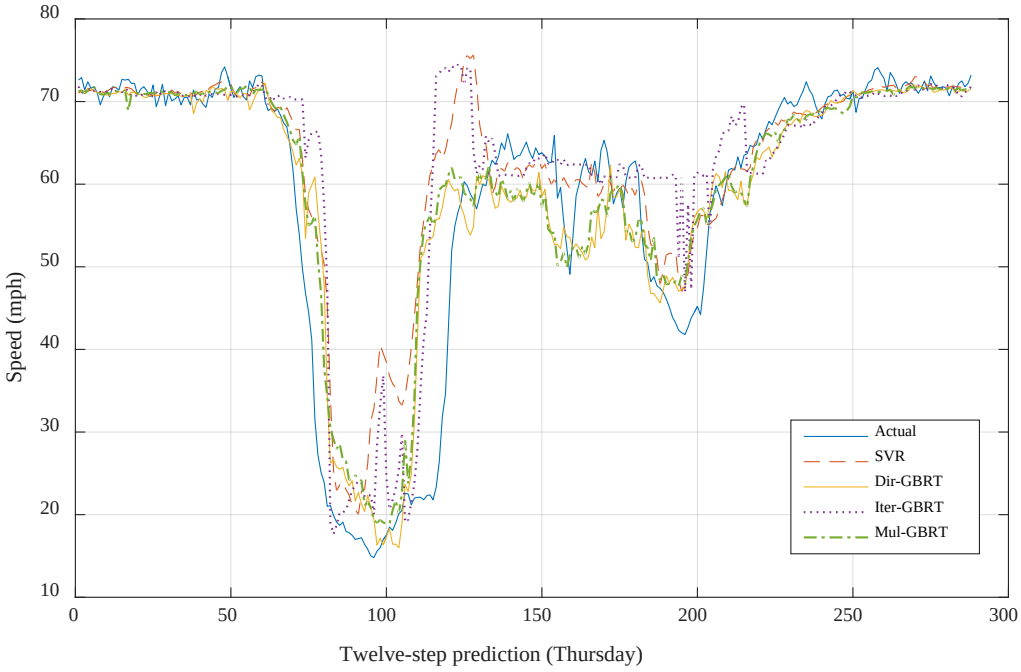
Model	30-min-ahead prediction		60-min-ahead prediction	
	Prediction time (s)	Rank	Prediction time (s)	Rank
SVR	3.26	4	6.57	4
Direct GBRT	0.96	2	1.94	2
Iterated GBRT	2.92	3	5.79	3
Multivariate GBRT	0.27	1	0.26	1

4.4.4 Prediction results

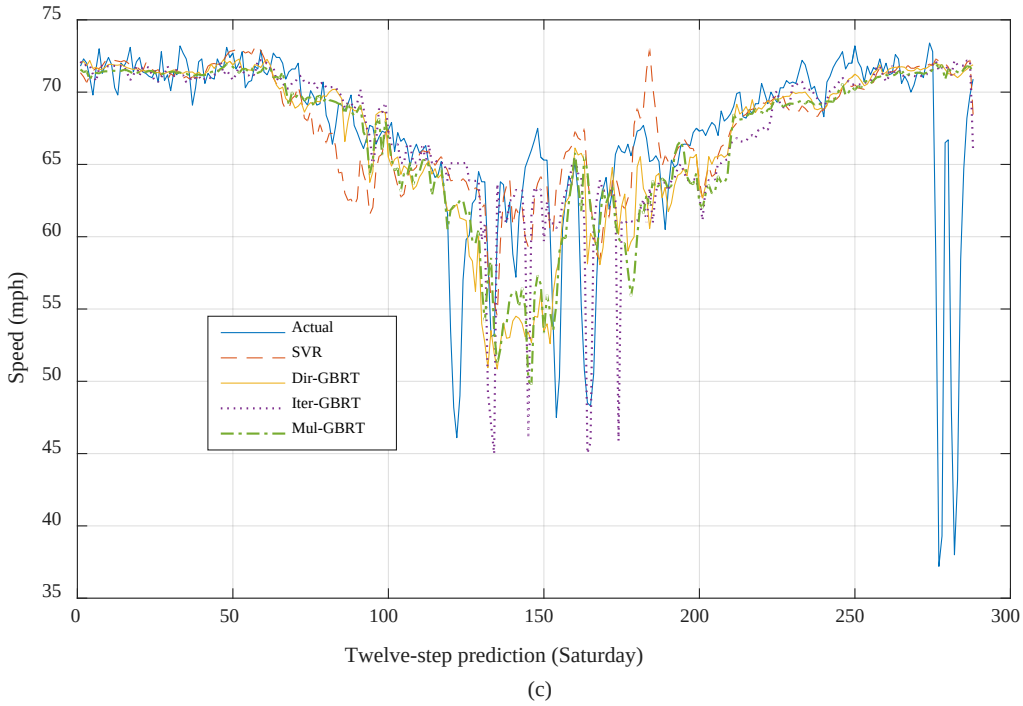
As shown in Figure 9, each model has good prediction performance in the first-step prediction, while the prediction accuracy obviously declines in the 12th prediction step. With the prediction step growth, the prediction ability decreases due to variable reasons, e.g., traffic congestion, accidents, morning and evening peaks. At the same time, the prediction ability on a working day is higher than that of the weekend for all models. Figure 9(b) shows that iterated GBRT and SVR have poor ability to predict the sudden change of speed caused by accidents.


Speed (mph)
80
70
60
50
40
30
20
10
0
50
100
150
200
250
300
Actual
SVR
Dir-GBRT
Iter-GBRT
Mul-GBRT
One-step prediction (Thursday)

(a)


Speed (mph)
80
70
60
50
40
30
20
10
0
50
100
150
200
250
300
Actual
SVR
Dir-GBRT
Iter-GBRT
Mul-GBRT
Twelve-step prediction (Thursday)

(b)



(c)

FIGURE 9 Comparison of prediction results on Thursday and Saturday.

4.4.5 Application in the scenario with target detector W4

To prove the applicability of the proposed model, we test it in another scenario, in which W4 is the target detector, W3 is the upstream detector, and W5 is the downstream detector. As shown in Table 13, we can conclude the same results as the scenario with target detector W3 we mentioned above. The multivariate GBRT has the advantage on the prediction stability and time. With the increasing prediction horizon, the multivariate GBRT performs better in prediction accuracy.

Table 13 Model performance comparison for target detector W4

	Model	Prediction accuracy (MAPE)	Prediction stability (MAPE)	Prediction time
30-min-ahead prediction	SVR	4.22%	**1.08%**	3.28
	Direct GBRT	**3.95%**	1.26%	0.97
	Iterated GBRT	3.99%	1.31%	2.93
	Multivariate GBRT	3.99%	1.26%	**0.24**
60-min-ahead prediction	SVR	5.05%	**1.14%**	6.56
	Direct GBRT	4.86%	1.27%	1.92
	Iterated GBRT	5.37%	1.73%	5.81
	Multivariate GBRT	**4.84%**	1.18%	**0.25**

5. CONCLUSIONS

In this paper, we conduct multi-step-ahead short-term traffic speed prediction by improving the gradient boosting regression tree. To solve multi-step-ahead prediction by GBRT, we use three strategies, namely direct GBRT, iterated GBRT, and multivariate GBRT. The proposed multivariate GBRT model achieves to output all prediction values simultaneously and consider the correlations between prediction variables through improving basic regression trees by taking into account all the predictor variables in the splitting process and modifying the splitting function. The models are evaluated by three criteria: prediction accuracy, prediction stability, and prediction time.

We explored the real-world data extracted from loop detectors in US101-N freeway through PeMS. SVR is used as the benchmark model and GBRT is taken as the basic model. The results show that: (I) Both direct GBRT and multivariate GBRT have obvious advantages in terms of the prediction accuracy compared to SVR; (II) Iterated GBRT has good performance in short-step prediction while the prediction accuracy decreases significantly with the increase of the prediction step due to the error accumulation; (III) Multivariate GBRT has the strongest prediction stability in multi-step-ahead prediction. With the prediction step growth, advantages of the high prediction stability are gradually reflected, which the prediction error distribution is relatively uniform; (IV) Multivariate GBRT outperforms the other models in the prediction time. Besides, with the increase of the prediction step, the advantages in prediction time are more obvious. Multivariate GBRT can save computational resources and improve the efficiency of prediction, which is significant for real-time applications.

In future research, we expect to learn how to select important features more efficiently and find a way to achieve dynamic calibration in the large-scale road networks. We will also compare the performance of random forest and GBRT in multi-step-ahead prediction.

ACKNOWLEDGMENTS

This research is financially supported by The National Key Research and Development Program of China [2018YFB1600904], National Natural Science Foundation of China [71771194, 71771198, 7181101222], Zhejiang Provincial Natural Science Foundation of China [LR17E080002], Key Research and Development Program of Zhejiang [2018C01007], and Young Elite Scientists Sponsorship Program by CAST [2018QNRC001].

REFERENCES

Ahmed, M., & Abdelaty, M. (2013). Application of stochastic gradient boosting technique to enhance reliability of real-time risk assessment. *Transportation Research Record: Journal of the Transportation Research Board*, 2386, 26-34.

Antoniou, C., Koutsopoulos, H. N., & Yannis, G. (2013). Dynamic data-driven local traffic state estimation and prediction. *Transportation Research Part C: Emerging Technologies*, 34(9), 89-107.

Bao, Y., Xiong, T., & Hu, Z. (2014a). Multi-step-ahead time series prediction using multiple-output support vector regression. *Neurocomputing*, 129(4), 482-493.

Bao, Y., Xiong, T., & Hu, Z. (2014b). PSO-mismo modeling strategy for multistep-ahead time series prediction. *IEEE Transactions on Cybernetics*, 44(5), 655-668.

Bontempi, G. (2008). Long term time series prediction with multi-input multi-output local learning. In *Proceedings of the 2nd European Symposium on Time Series Prediction (TSP), ESTSP08*, Helsinki, Finland, 2008, pp. 145-154.

Ben Taieb, S., Sorjamaa, A., & Bontempi, G. (2010). Multiple-output modeling for multi-step-ahead time series forecasting. *Neurocomputing*, 73(10), 1950-1957.

Ben Taieb, S., Bontempi, G., Atiya, A.F., & Sorjamaa, A. (2012). A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *Expert Systems with Applications*. 39, 7067-7083.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Chevillon, G. (2007). Direct multi-step estimation and forecasting. *Journal of Economic Surveys*, 21(4), 746-785.

Chung, Y. S. (2013). Factor complexity of crash occurrence: An empirical demonstration using boosted regression trees. *Accident Analysis & Prevention*, 61(8), 107-118.

Cox, D. R. (1961). Prediction by exponentially weighted moving averages and related methods. *Journal of the Royal Statistical Society*, 23(2), 414-422.

Dumont, M., Marée, R., Wehenkel, L., & Geurts, P. (2009). Fast multi-class image annotation with random subwindows and multiple output randomized trees. In *Proceedings of the Fourth International Conference on Computer Vision Theory and Applications*, Lisbon, Portugal, Vol. 2, pp. 196-203.

Farokhi Sadabadi, K., Hamedi, M., & Haghani, A. (2010). Evaluating moving average techniques in short-term travel time prediction using an AVI data set. In *Transportation Research Board 89th Annual Meeting*, Washington, DC.

Fei, X., Lu, C. C., & Liu, K. (2011). A Bayesian dynamic linear model approach for real-time short-term freeway travel time prediction. *Transportation Research Part C: Emerging Technologies*, 19(6), 1306-1318.

Guo, F., Krishnan, R., & Polak, J. (2017). The influence of alternative data smoothing prediction techniques on the performance of a two-stage short-term urban travel time prediction framework. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, 21(3), 214-226.

Hamner, B. (2010). Predicting travel times with context-dependent random forests by modeling local and aggregate traffic flow. In *IEEE International Conference on Data Mining Workshops*, Sydney, Australia, pp. 1357-1359.

Hinsbergen, C. P. I. V., Lint, J. W. C. V., & Zuylen, H. J. V. (2009). Bayesian committee of neural networks to predict travel times with confidence intervals. *Transportation Research Part C: Emerging Technologies*, 17(5), 498-509.

Hu, Z., Bao, Y., & Xiong, T. (2014). *Comprehensive learning particle swarm optimization based memetic algorithm for model selection in short-term load forecasting using support vector regression*. Applied Soft Computing, 25(C), 15-25.

Ikeguchi, T., & Aihara, K. (1995). Prediction of chaotic time series with noise. *IEICE Transactions on Fundamentals of Electronics Communications & Computer Sciences*, 78(10), 1291-1298.

Leshem, G. (2007). Traffic flow prediction using AdaBoost algorithm with random forests as a weak learner. *Enformatika*, 193, 193-198.

Li, L., Chen, X., Li, Z., & Zhang, L. (2013). Freeway travel-time estimation based on temporal-spatial queueing model. *IEEE Transactions on Intelligent Transportation Systems*, 14(3), 1536-1541.

Li, L., Chen, X., & Zhang, L. (2014). Multimodel ensemble for freeway traffic state estimations. *IEEE Transactions on Intelligent Transportation Systems*, 15(3), 1323-1336.

Min, W., & Wynter, L. (2011). Real-time road traffic prediction with spatio-temporal correlations. *Transportation Research Part C: Emerging Technologies*, 19(4), 606-616.

Qiao, W., Haghani, A., Shao, C. F., & Liu, J. (2016). Freeway path travel time prediction based on heterogeneous traffic data through nonparametric model. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, 20(5), 438-448.

Segal, M., & Xiao, Y. (2011). Multivariate random forests. *Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery*, 1(1), 80–87.

Segal, M. R. (1992). Tree-structured methods for longitudinal data. *Journal of the American Statistical Association*, 87(418), 407-418.

Sorjamaa, A., Hao, J., Reyhani, N., Ji, Y., & Lendasse, A. (2007). Methodology for long-term prediction of time series. *Neurocomputing*, 70(16), 2861-2869.

Vlahogianni, E. I., Karlaftis, M. G., & Golias, J. C. (2014). Short-term traffic forecasting: where we are and where we're going. *Transportation Research Part C: Emerging Technologies*, 43, 3-19.

Wang, J., & Shi, Q. (2013). Short-term traffic speed forecasting hybrid model based on chaos–wavelet analysis-support vector machine theory. *Transportation Research Part C: Emerging Technologies*, 27(2), 219-232.

Wang, Y. (2011). Prediction of weather impacted airport capacity using ensemble learning. Digital Avionics Systems Conference (Vol.49, pp. 2D6-1-2D6-11). IEEE.

Wei, Y., & Chen, M. C. (2012). Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks. *Transportation Research Part C: Emerging Technologies*, 21(1), 148-162.

Williams, B., Durvasula, P., & Brown, D. (1998). Urban freeway traffic flow prediction: application of seasonal autoregressive integrated moving average and exponential smoothing models. *Transportation Research Record: Journal of the Transportation Research Board*, 1644(1), 132-141.

Williams, R. J., & Zipser, D. (2014). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2), 270-280.

Wu, Y. J., Chen, F., Lu, C. T., & Yang, S. (2016). Urban traffic flow prediction using a spatio-temporal random effects model. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, 20(3), 282-293.

Xiong, T., Bao, Y., & Hu, Z. (2013). Beyond one-step-ahead forecasting: evaluation of alternative multi-step-ahead forecasting models for crude oil prices. *Energy Economics*, 40(2), 405-415.

Xiong, T., Bao, Y., & Hu, Z. (2014). Does restraining end effect matter in EMD-based modeling framework for time series prediction? Some experimental evidences. *Neurocomputing*, 123, 174-184.

Zhang, F., Zhu, X., Hu, T., Guo, W., Chen, C., & Liu, L. (2016). Urban link travel time prediction based on a gradient boosting method considering spatiotemporal correlations. *ISPRS International Journal of Geo-Information*, 5(11), 201.

Zhang, X. & Rice, J. A. (2003). Short-term travel time prediction. *Transportation Research Part C: Emerging Technologies*, 11(3), 187-210.

Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, 58, 308-324.