IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 13, 2020, accepted February 24, 2020, date of publication February 28, 2020, date of current version March 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2977219

A Hybrid Model for Lane-Level Traffic Flow Forecasting Based on Complete Ensemble Empirical Mode Decomposition and Extreme Gradient Boosting

WENQI LU[1,2,3], YIKANG RUI[1,2,3], ZIWEI YI[1,2,3], BIN RAN[1,2,3], AND YUANLI GU[4]

[1]School of Transportation, Southeast University, Nanjing 211189, China
[2]Joint Research Institute on Internet of Mobility, Southeast University and University of Wisconsin-Madison, Southeast University, Nanjing 211189, China
[3]Jiangsu Key Laboratory of Urban ITS, Southeast University, Nanjing 211189, China
[4]Key Laboratory of Transport Industry of Big Data Application Technologies for Comprehensive Transport, Ministry of Transport, Beijing Jiaotong University, Beijing 100044, China

Corresponding author: Yikang Rui (101012189@seu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 41971342, in part by the Science and Technology Program of Beijing under Grant Z121100000312101, and in part by the Fundamental Research Funds for the Central Universities under Grant 2242019k30054.

ABSTRACT Accurate and efficient lane-level traffic flow prediction is a challenging issue in the framework of the connected automated vehicle highway system. However, most existing traffic flow forecasting methods concentrate on mining the spatio-temporal characteristics of the traffic flow rather than increasing predictability of traffic flow. In this paper, we propose a novel hybrid model (CEEMDAN-XGBoost) for lane-level traffic flow prediction based on complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) and extreme gradient boosting (XGBoost). The CEEMDAN method is introduced to decompose the raw traffic flow data into several intrinsic mode function components and one residual component. Then, the XGBoost methods are trained and make predictions on the decomposed components respectively. The final prediction results are obtained by integrating the prediction outputs of the XGBoost methods. For illustrative purposes, the ground-truth lane-level traffic flow data captured by remote traffic microwave sensors installed on the 3rd Ring Road of Beijing are utilized to evaluate the effectiveness of the CEEMDAN-XGBoost model. The experimental results confirm that the CEEMDAN-XGBoost model is capable of fitting the complex volatility of traffic flow efficiently at different types of lane sections. Moreover, the proposed model outperforms the state-of-the-art models (e.g., artificial neural networks and long short-term memory neural network) and other XGBoost-based models in terms of prediction accuracy and stability.

INDEX TERMS Data mining, lane-level traffic flow, short-term prediction, hybrid model, extreme gradient boosting, complete ensemble empirical mode decomposition, urban expressways.

I. INTRODUCTION

Accurate and timely traffic flow forecasting is critical for the successful development of intelligent transportation systems (ITS). It can benefit both traffic management agencies and travelers by contributing to various kinds of key applications such as variable speed limit control and route guidance systems. During the past three decades, the combination of unprecedented data availability and the ability to process these data has brought on immense development and spread of ITS technologies [1]. At the same time, a novel data-driven research area has been systematically growing in parallel to the ell-founded mathematical models that are based on macroscopic and microscopic theories of traffic flow [2].

However, with the rapid improvement of the connected automated vehicle highway (CAVH) system [3], short-term traffic flow forecasting has been gradually shifting from the section-based or network-based methods to lane-based methods [4]. In the environment of the CAVH system, the traffic flow on the road is generally mixed with human-driven

The associate editor coordinating the review of this manuscript and approving it for publication was Huiling Chen.

vehicles (HDVs) and connected automated vehicles (CAVs) [5]. Forecasting the traffic flow of the lane sections in a future period which varies from several minutes to dozens of minutes is necessary for both CAVs and HDVs. On the one hand, the lane-level prediction of dynamic traffic flow can provide more real-time and detailed traffic state information for the HDVs to choose the appropriate travel route and overcome the influence of the limited sight distance [6]. On the other hand, high-efficiency lane-level prediction can assist the CAVs in making lane selection and planning the optimal travel trajectory in terms of the level of service based on the predicted traffic flow [7]. Sequentially, the overall distribution of various vehicles on the roads can be more balanced and the road capacity will be improved correspondingly.

Although many methods for short-term traffic flow prediction have been proposed during the past several decades, some limitations and challenges still exist as follows. 1) Most existing studies forecast aggregated traffic flow rather than lane-level traffic flow with the implicit hypothesis that traffic patterns between different lanes are the same or exactly similar. The traffic flows are often aggregated for simplifying the model complexity or lacking lane-level traffic data. However, some studies have found that the traffic flows of different lanes indicate different patterns[8]–[10]. Furthermore, compared with the traffic flow at the road sections, the volatility and uncertainty of the traffic flow at lane sections are more significant due to the lane failure phenomenon [11] and lane-drop bottleneck [12], which increases the difficulty of making lane-level traffic flow prediction. 2) Most existing methods attempt to improve the performance of the prediction model by considering spatio-temporal dependence and correlation of traffic flow rather than focusing on improving the predictability of traffic flow waveform itself. Meanwhile, though the traffic flow prediction methods [13]–[15] based on advanced deep learning approaches have the advantages of capturing the complex characteristics of the traffic flow, the training time consumption of deep learning-based methods are still too long to satisfy the real-time requirements of the CAVs and CAVH.

To overcome the aforementioned problems, this paper intends to put forward a novel hybrid model named CEEMDAN-XGBoost for forecasting lane-level traffic flow through increasing the predictability of the complex lane-based traffic flow. We exploit the complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) to decompose the complicated and irregular traffic flow into several low-noise components. Then, an improved boosting method named extreme gradient boosting (XGBoost) is chosen as the predictor in the framework of the proposed model. The XGBoost predictors are trained and make predictions on each component. By combining the prediction values of all XGBoost predictors, the proposed model can make full use of the temporal feature of traffic flow and obtain the precise prediction results in this way.

In summary, the major contributions of this paper are presented as follows:

- In this study, the raw traffic flow data of the lane sections are selected as the research target instead of those of road sections or large-scale road network.
- To the best of our knowledge, a valid signal processing method named complete ensemble empirical mode decomposition with adaptive noise is first introduced to produce the predictable and regular decomposed components of the raw traffic flow, which can decrease the unpredictability of the lane-level traffic flow.
- A novel hybrid model named CEEMDAN-XGBoost, which fuses CEEMDAN and XGBoost algorithm effectively, is established to realize both single-step-ahead and multi-step-ahead lane-level traffic flow prediction.
- Validated by real-world traffic flow data of lanes captured by several remote traffic microwave sensors (RTMS) installed on the 3rd Ring Road of Beijing with the sampling time interval of 2 min, the CEEMDAN-XGBoost model outperforms both the traditional and state-of-the-art benchmark models in terms of prediction accuracy and stability.

The remainder of this paper is organized as follows. A general overview of existing literature on traffic forecasting is provided in section II. Section III gives a detailed description of the CEEMDAN method, XGBoost algorithm, and the proposed hybrid method. Section IV introduces the experimental dataset, environment, and evaluation index. Section V discusses the experimental results and analysis. At last, the conclusion and future work are given in section VI.

II. RELATED WORK

On the whole, the existing traffic flow prediction models can be categorized into the following three groups from the perspective of methodology: traditional parametric models, artificial intelligent-based models, and hybrid models.

A. TRADITIONAL PARAMETRIC MODELS

The classical parametric approaches for traffic flow prediction mainly include the Kalman filtering methods [16], exponential smoothing methods [17], auto-regressive integrated moving average (ARIMA) models [18], the structural time-series models [19], and multivariate time series models [20]. In the beginning, the parametric models such as ARIMA and ARIMA-based models [21]–[23] illustrated great performance in terms of making traffic prediction in the short term. Hamed *et al.* [24] applied an ARIMA model to forecast traffic volume prediction in urban arterial roads. Nevertheless, most of these parametric models cannot process the complex patterns in traffic flow because they are built with some presumption [25]. To solve the above-mentioned problem, relevant scholars put forward many artificial intelligent-based models [26] whose structures and parameters are more flexible.

B. ARTIFICIAL INTELLIGENT-BASED MODELS

The common artificial intelligent-based models contain support vector machines [27], fuzzy logic system methods [28],

k-nearest neighbour [29]–[31] and artificial neural networks (ANN) [32]–[34]. Among these models, ANNs are considered as another popular countermeasure for traffic prediction due to their capability of handling multi-dimensional data and mining the complex patterns of the measured historical data [2]. Vlahogianni *et al.* [35] put forward a neural predictor, which was composed of time-optimized multilayer perceptron (MLP) structures, to provide accurate short-term traffic flow prediction by using spatio-temporal data. Tang *et al.* [36] proposed an improved fuzzy neural network for multi-step traffic speed forecasting by considering the periodic characteristic of the traffic flow. With the remarkable improvements of data storage and processing technology, the popular short-term traffic prediction methods have shifted from ANN-based methods to deep learning methods which can automatically discover the implicit relationships inside the data using a general-purpose learning procedure. Ma *et al.* [37] firstly introduced long short-term memory (LSTM) neural network, which can capture the long temporal dependency for the input sequence, to predict the traffic speed data of the expressways in Beijing. Chen *et al.* [38] proposed an improved LSTM network considering spatio-temporal correlation in traffic system via a two-dimensional network to achieve better prediction performance. Yang *et al.* [39] built a stacked auto-encoder Levenberg-Marquardt model to improve forecasting accuracy. Though the deep learning-based models can learn the spatio-temporal characteristics efficiently, their prediction performance heavily depend on the quantity and quality of the dataset. Therefore, single models, even the deep learning models, are still difficult in dealing with the dynamic fluctuations of the traffic flow.

C. HYBRID MODELS

To address the aforementioned problems, hybrid models [40]–[43] may be a better choice to solve the traffic prediction problem. Vlahogianni [44] proposed a surrogate model considering fusing three different models to forecast the short-term speed on the freeway. Li *et al.* [45] put forward a deep belief network optimized by the multi-objective particle swarm algorithm to realize multi-time-step forecasting. Wu *et al.* [14] established a hybrid deep neural network, which employs a convolutional neural network to mine the spatial features and uses the recurrent neural network to mine the temporal features of traffic flow, to predict traffic flow in a long-term horizon. Gu *et al.* [4] put forward an improved Bayesian combination method which fuses a traditional parametric model, a non-parametric model, and an RNN-based model to take advantage of each method. Wang *et al.* [46] combined the empirical mode decomposition (EMD) with the ARIMA model to predict traffic speeds in varying scenarios such as mixed traffic flow. Based on Wang's study, Li *et al.* [47] fused an ensemble EMD and random vector functional link network to predict travel time. Wei and Chen [48] developed a hybrid model combining the EMD model and back-propagation neural networks (BPNN) to forecast the short-term metro passenger flow. Therefore, the thought of decomposing the raw traffic flow data to obtain more predictable components has proved to be an efficient way to improve the performance of normal methods.

Hence, this paper aims to establish a decomposition-based hybrid model for lane-level traffic flow prediction. In the proposed model, a novel signal decomposition method named CEEMDAN is introduced to deal with raw traffic flow data so as to improve the predictability of sensitive lane-level traffic flow. Then, the XGBoost predictors are trained and employed to make predictions on the decomposed components respectively. Finally, the predicted values of all components are integrated to obtain the forecasted traffic flow.

III. METHODOLOGY

In this section, the theoretical backgrounds of the CEEMDAN approach and the XGBoost method are introduced. Then, a detailed description of the CEEMDAN-XGBoost model is presented.

A. COMPLETE ENSEMBLE EMPIRICAL MODE DECOMPOSITION WITH ADAPTIVE NOISE

CEEMDAN is a variation of ensemble empirical mode decomposition (EEMD) algorithm [49], which provides an exact reconstruction of the original signal and a better spectral separation of the modes with a lower computational cost. Hence, we first introduce the EEMD method and then extend it to the CEEMDAN.

EEMD is an efficient noise-assisted data analysis method based on EMD [50]. It can overcome the mode mixing problem of the EMD by adding white Gaussian noise to the raw data [51]. The EMD approach decomposes the raw data into several intrinsic mode functions (IMF) or modes. Note that a signal considered as an IMF must satisfy the two following conditions. First, the number of extrema points and the number of zero-crossing points must be equal to 1 or differ at most by 1. Second, the mean value of the upper and lower envelope is 0 everywhere.

EEMD defines the final IMF components as the mean of the corresponding IMFs obtained via EMD over an ensemble of trials, generated by adding different realizations of the white noise of finite variance to the original data $Y(n)$. Note that the generated IMF components are defined as $\overline{IMF}_m(n)$, $m = 1, 2, \ldots, M$ and M is the number of the IMF components. The process of implementing the EEMD algorithm can be described in the following steps:

Step 1: Add different realizations of white Gaussian noise $\omega^k(n), k = 1, 2, \ldots, K$ to the raw data $Y(n)$ and generate the noise-added data $Y^k(n) = Y(n) + \omega^k(n), k = 1, 2, \ldots, K$, where K is the number of realizations.

Step 2: Each $Y^k(n), k = 1, 2, \ldots, K$ is decomposed by using the EMD algorithm to achieve the mode $IMF_m^k(n)$, where $IMF_m^k(n), m = 1, 2, \ldots, M$ refers to the m-th IMF mode of $Y^k(n), k = 1, 2, \ldots, K$ and M is the number of modes.

Step 3: Calculate the m-th IMF mode of $Y(n)$, which is obtained as the average of the corresponding $IMF_m^k(n)$ and is defined as follows:

$$\overline{IMF}_m(n) = \frac{1}{K}\sum_{k=1}^{K} IMF_m^k(n) \tag{1}$$

Though the EEMD algorithm is able to overcome the mode mixing problem by populating the whole time-frequency space to take advantage of the dyadic filter bank behaviour of the EMD algorithm. Some weakness still exists. For instance, the reconstructed signal includes residual noise and different realizations of signal plus noise may produce a different number of modes [49].

To address the above problem and improve the effect of decomposition, Torres *et al.* [49] proposed a novel signal decomposition approach named CEEMDAN. It has been illustrated in the EEMD algorithm that each $Y^k(n)$ is decomposed independently from the other realizations and a residue $\tau_m^k(n)$ for each one can be obtained as follows:

$$\tau_m^k(n) = \tau_{m-1}^k(n) - IMF_m^k(n) \tag{2}$$

In the CEEMDAN method, the decomposition modes are donated as $\widehat{IMF}_m(n)$ and the first residue can be formulated as:

$$\tau_1(n) = Y(n) - \widehat{IMF}_1(n) \tag{3}$$

where $\widehat{IMF}_1(n)$ is achieved by utilizing the same way in the EEMD.

Then, the first EMD mode can be computed over an ensemble of $\tau_1(n)$ plus different realizations of a given noise obtaining $\widehat{IMF}_2(n)$ by averaging. The next residue can be written as:

$$\tau_2(n) = \tau_1(n) - \widehat{IMF}_2(n) \tag{4}$$

This procedure continues with the rest of the modes until the stopping criterion is satisfied. Let define the function E_m which produces the m-th mode obtained by using EMD algorithm and define the function $Num(\cdot)$ which generates the number of extreme points. ω^k is the white noise with a distribution of $N(0,1)$ and ε_k is the noise standard deviation. Therefore, if $Y(n)$ is the raw data, the pseudo-code of implementing the CEEMDAN algorithm is described in **Algorithm 1**. The given raw data $Y(n)$ can be expressed in the following equation:

$$Y(n) = \sum_{m=1}^{M} \widehat{IMF}_m + r(n) \tag{5}$$

Note that the coefficients ε_k allow selecting the signal-noise ratio (SNR) at each stage of the decomposition. According to related studies [49], [52], a few hundred of realizations and the same SNR for all the stages can be employed in the CEEMDAN algorithm.

Algorithm 1 The Realization of the CEEMDAN Approach

Input: $Y(n)$—The raw time-series data
K— The number of realizations
Output: $\widehat{IMF}_m(n), m = 1, 2, \ldots, M$ and $r(n)$

for $k = 1$ to K **do**
  $IMF_1^k(n) \leftarrow E_1(Y(n) + \varepsilon_0\omega^k(n))$
end
$\widehat{IMF}_1(n) \leftarrow \left(\sum_{k=1}^{K} IMF_1^k(n)\right) \Big/ K$
$\tau_1(n) \leftarrow Y(n) - \widehat{IMF}_1(n)$
$m = 1$
while $Num(\tau_m(n)) \geq 2$ **do**
  for $k = 1$ to K **do**
    $IMF_{m+1}^k(n) \leftarrow E_1(\tau_m(n) + \varepsilon_m E_m(\omega^k(n)))$
  end
  $\widehat{IMF}_{m+1}(n) \leftarrow \left(\sum_{k=1}^{K} IMF_{m+1}^k(n)\right) \Big/ K$
  $\tau_{m+1}(n) \leftarrow \tau_m(n) - \widehat{IMF}_{m+1}(n)$
  $m \leftarrow m + 1$
end
$M \leftarrow m - 1$
$r(n) \leftarrow Y(n) - \sum_{m=1}^{M} \widehat{IMF}_m$

B. EXTREME GRADIENT BOOSTING ALGORITHM

Extreme gradient boosting (XGBoost) [53] is a scalable and portable boosting algorithm in ensemble learning, which follows the gradient boosting framework proposed by Friedman [54]. XGBoost has gained great popularity and attention since its release due to its excellent performance on a broad range of machine learning competitions [53]. XGBoost considers the regularization term and the target function is defined as follows:

$$L(\theta) = \sum_i l(\hat{y}_i, y_i) + \sum_k \psi(f_k) \tag{6}$$

$$\psi(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{7}$$

where $\hat{y}_i$ is the predicted value; y_i is the ground-truth value; l is a differentiable convex loss function. f_k corresponds to an independent tree structure q and leaf weights w; ψ penalizes the complexity of the model; T represents the number of leaves in the tree; λ is donated as the punishing regularization term helping to smooth the final learnt weights to solve the over-fitting problem; γ stands for the punishing regularization of the leaf tree which has pruning function.

The tree ensemble model in (6) and (7) contains the functions as parameters which cannot be solved through using traditional optimization methods. Therefore, another way is used to train the model. Donate $\hat{y}_i^{(t)}$ as the predicted value of the i-th data at the t-th step. Another tree f_t is added to

minimize the objective as follows:

$$L^{(t)} = \sum_{i=1}^{N} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \psi(f_t) \tag{8}$$

Taylor formula can be utilized to transform the objective function (8) into a second-order polynomial [55], which is given by:

$$\begin{aligned}\tilde{L}^{(t)} &= \sum_{i=1}^{N} [l(y_i, \hat{y}_i^{(t-1)}) + h_i' f_t(x_i) + \frac{1}{2} h_i'' f_t(x_i)] + \psi(f_t) \\ &= \sum_{i=1}^{N} [h_i' f_t(x_i) + \frac{1}{2} h_i'' f_t^2(x_i)] + \psi(f_t) + c \end{aligned} \tag{9}$$

where $h_i' = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ and $h_i'' = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ are the first-order and the second-order gradient statistics on the loss function respectively; c represents the constant item.

By removing the constant terms, (9) can be simplified as:

$$\tilde{L}^{(t)} = \sum_{i=1}^{N} [h_i' f_t(x_i) + \frac{1}{2} h_i'' f_t^2(x_i)] + \psi(f_t) \tag{10}$$

Set $Q_j = \{i | q(x_i) = j\}$ as the instance set of the j-th leaf node and model objective function is calculated as:

$$\begin{aligned}\tilde{L}^{(t)} &= \sum_{i=1}^{N} [h_i' f_t(x_i) + \frac{1}{2} h_i'' f_t^2(x_i)] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} [(\sum_{i \in I_j} h_i') w_j + \frac{1}{2} (\sum_{i \in I_j} h_i'' + \lambda) w_j^2] + \gamma T \end{aligned} \tag{11}$$

For a fixed tree structure $q(x)$, the optimal weight w_j^* of the j-th leaf node and the corresponding optimal objective function values can be respectively expressed as:

$$w_j^* = -\frac{\sum_{i \in I_j} h_i'}{\sum_{i \in I_j} h_i'' + \lambda} \tag{12}$$

$$\tilde{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} h_i')^2}{\sum_{i \in I_j} h_i'' + \lambda} + \gamma T \tag{13}$$

Equation (13) can be utilized to measure the quality of the tree structure. However, it is impossible to iterate over all possible structures of the trees. Hence, a greedy algorithm which starts with a leaf and iteratively adds branches to the tree is introduced to look for the optimal tree structure. In practice, the evaluation index for the split candidates is formulated as:

$$\begin{aligned} &L_{split} \\ &= -\frac{1}{2} \left[\frac{(\sum_{i \in I_L} h_i')^2}{\sum_{i \in I_L} h_i'' + \lambda} + \frac{(\sum_{i \in I_R} h_i')^2}{\sum_{i \in I_R} h_i'' + \lambda} - \frac{(\sum_{i \in I} h_i')^2}{\sum_{i \in I} h_i'' + \lambda} \right] - \gamma \end{aligned} \tag{14}$$

where I_L and I_R are the instance sets of left and right nodes after splitting; $I = I_L \cup I_R$.

Algorithm 2 Sparsity-Aware Split Finding

Input: I—Instance set of the current node
$I_k = \{i \in I | x_{ik} \neq \text{missing}\}$
d—Feature dimension
Output: Split and default directions with max *gain*

```
gain ← 0   G ← Σ_{i∈I} h_i'   H ← Σ_{i∈I} h_i''
for k = 1 to m do
    G_L ← 0, H_L ← 0
    for j in sorted (I_k, ascent order by x_ik) do
        G_L ← G_L + h_j', H_L ← H_L + h_j''
        G_R ← G − G_L, H_R ← H − H_L
        G_b ← G_L^2/(H_L + λ) + G_R^2/(H_R + λ) − G^2/(H + λ)
        score ← max(score, G_b)
    end
    G_R ← 0, H_R ← 0
    for j in sorted (I_k, ascent order by x_ik) do
        G_R ← G_R + h_j', H_R ← H_R + h_j''
        G_L ← G − G_R, H_L ← H − H_R
        G_b ← G_L^2/(H_L + λ) + G_R^2/(H_R + λ) − G^2/(H + λ)
        score ← max(score, G_b)
    end
end
```

However, a challenging problem is to find the optimal splitting in (14) because it is difficult to compute all possible splits of continuous features. As the input $\boldsymbol{x}$ of the short-term traffic flow prediction is sparse, a sparsity-aware split finding algorithm proposed by Chen and Guestrin [53] is exploited to look for the optimal splitting in this paper. **Algorithm 2** shows the pseudo-code of the sparsity-aware split finding algorithm. It illustrates that the instance is classified into the default direction when a value is missing in the sparse matrix $\boldsymbol{x}$. The optimal default directions are learnt from the data.

C. A HYBRID MODEL BASED ON CEEMDAN AND XGBOOST

To improve the predictability of lane-level traffic flow data and forecast lane-level traffic flow more efficiently, this paper establishes a novel decomposition-based hybrid model named CEEMDAN-XGBoost. The proposed hybrid model combines the CEEMDAN method and the XGBoost method. This section introduces the overall architecture of the proposed model revealed in Figure.1 and illustrates how to utilize it to implement lane-level traffic flow prediction.

Firstly, the original traffic flow data of the target lane section can be captured with a sampling interval by detectors. The raw traffic flow data (volume or speed) of the lane can be donated as $Y(n)$.

Secondly, the CEEMDAN method is introduced to decompose the original traffic flow data $Y(n)$ into several IMF

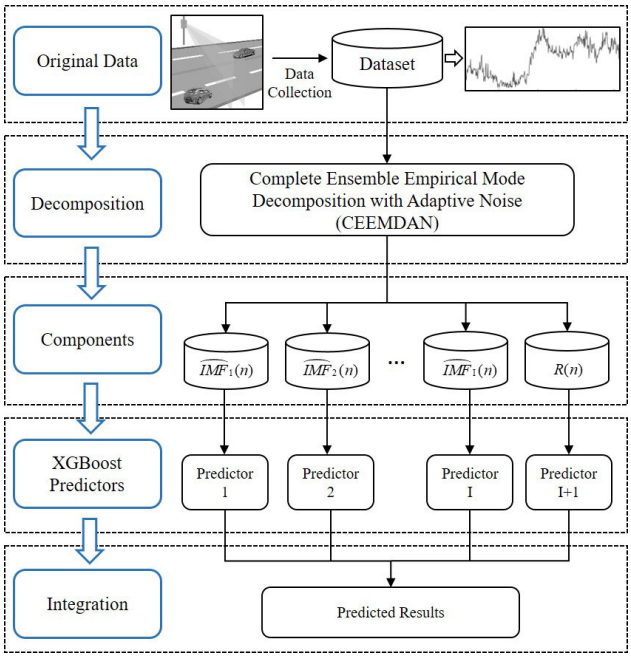


FIGURE 1. The architecture of the CEEMDAN-XGBoost model.

components $\widehat{IMF}_i(n), i = 1, 2, \ldots, I$ and a residual component $R(n)$. I is the number of IMF components. The specific decomposition algorithm of the CEEMDAN is shown in **Algorithm 1**. The CEEMDAN method enables to transform the complicated raw traffic flow data of the lane section into several regular de-noised components, which are beneficial to reduce the difficulty and complexity of predicting the lane-level short-term traffic flow.

Then, the XGBoost predictors can be trained in parallel based on the historical time series data of the decomposed components by using **Algorithm 2**. Let $q_i(t)$ donate the element of the $\widehat{IMF}_i(n)$ and $R(n)$ at the time interval t. For all XGBoost predictors, the input time step and the maximum depth of a tree of each component are set as h_s and m respectively. Hence, the input of the i-th component at the time interval t can be defined as $Q_i(t), i = 1, 2, \ldots, I + 1$, which is written as:

$$Q_i(t) = [q_i(t - h_s), q_i(t - h_s + 1), \ldots, q_i(t - 1)]^T \quad (15)$$

Meanwhile, during the training process of the XGBoost predictors, the training output of the i-th component at the time interval t is defined as $q_i(t)$. Based on the above discussion, the implementation of training a CEEMDAN-XGBoost model can be divided into two phases illustrated in **Algorithm 3**.

Afterwards, the trained XGBoost predictors are exploited to make predictions on the decomposed components separately at each prediction time interval. The prediction output of the i-th XGBoost predictor at the time interval t donated as $\widehat{q}_i(t)$ can be obtained.

Algorithm 3 Realization of Training a CEEMDAN-XGBoost Model

Input: $Q(n)$—The raw time series data of traffic flow
h_s—The input time step
m—The maximum depth of a tree
Output: The trained CEEMDAN-XGBoost model

```
Phase I: Raw traffic flow decomposition
Decompose Q(n) into _i(n) and R(n) by Algorithm 1
Phase II: Training process of the XGBoost predictors
for i = 1 to I + 1 do
    T_in ← [Q_i(h_s + 1), Q_i(h_s + 2), ..., Q_i(n)]^T
    T_out ← [q_i(h_s + 1), q_i(h_s + 2), ..., q_i(n)]^T
    Train i − th XGBoost Predictor with T_in and T_out by
    Algorithm 2
end
return trained CEEMDAN-XGBoost model
```

At last, the prediction traffic flow data of the lane section at the time interval t is the sum of the prediction values of all XGBoost predictors in the CEEMDAN-XGBoost model. The final fusion prediction result is written as $\bar{q}(t)$, which can be calculated as follows:

$$\bar{q}(t) = \sum_{i=1}^{I+1} \widehat{q}_i(t) \quad (16)$$

On the whole, the procedures for implementing lane-based traffic flow prediction by utilizing the CEEMDAN-XGBoost model are summarized as follows:

Step 1: Collect the raw traffic flow data of the target lane section by detectors.

Step 2: Employ the CEEMDAN method to decompose the raw data into several IMF components and a residual component, and divide the decomposed time series data into a training dataset and a testing dataset.

Step 3: Train the XGBoost predictors in parallel with the training dataset by setting relevant parameters.

Step 4: Predict the values of the components by using the trained XGBoost predictors at each prediction time interval.

Step 5: Calculate the predicted traffic flow of the lane section at each prediction time intervals by accumulating the predicted values of all XGBoost predictors.

IV. EXPERIMENT

A. EXPERIMENTAL DATASET AND ENVIRONMENT

To test the practicability and accuracy of the proposed model, the real-world traffic flow data of lanes captured by the remote traffic microwave sensors (RTMS) located at six road sections of the 3rd Ring Road in Beijing were employed to carry out relevant experiments. TABLE 1 and Figure.2 provide observation locations and detailed information about the lane sections. The sampling periods of the traffic flow range from 2014.1.6 to 2014.1.19 and from 2014.2.17 to 2014.2.23 with the sampling time interval of 2 min. Hence, the entire length and the total number of the records at each

TABLE 1. The detailed information of the observation lane sections.

Road ID	Name of Road section	Number of lanes	Lane section ID
P1	The South of Yansha Bridge	3	L1,L2,L3
P2	The South of Changhong Bridge	3	L11,L12,L13
P3	The North of Jinsong Bridge	3	L1,L2,L3
P4	The East of Fenzhongsi Bridge	3	L11,L12,L13
P5	The West of Fangzhuang Bridge	3	L1,L2,L3
P6	The West of Dongtieying Bridge	3	L11,L12,L13

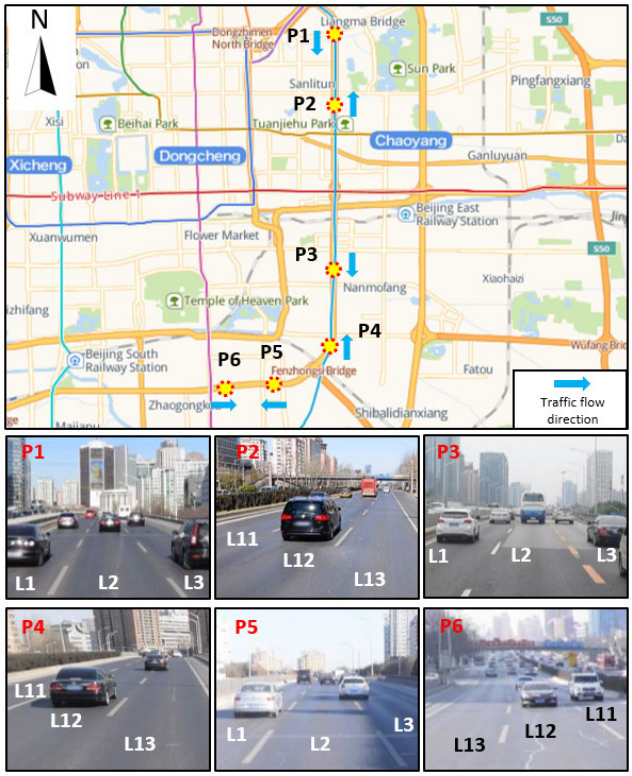


FIGURE 2. The location of the observation lane sections.

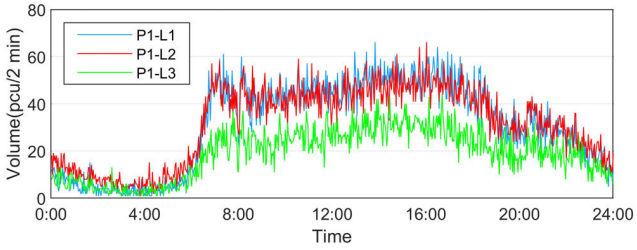


FIGURE 3. The example of raw traffic flow of at road section P1 during a day.

lane section are three weeks and 15120 with data validity rate higher than 95% [56].

To ensure the quality of the raw traffic flow data and improve the reliability of the prediction results, we took the threshold approach and traffic flow rule-based approach to identify and eliminate abnormal data. Then, the missing and erroneous records were properly remedied by using temporally adjacent records [37]. Figure.3 reveals the raw traffic flow data during a workday at three lanes of the road section P1.

Note that the entire dataset was divided into three parts. The first part (2014.1.6-2014.1.10 and 2014.1.13-2014.1.17) and the second part (2014.1.11-2014.1.12 and 2014.1.18-2014.1.19) were used to train the parameters of the models. The two parts contained the data during workdays and weekends respectively. The rest part (2014.2.17-2014.2.23) of the dataset was employed to test the models with trained parameters.

The experimental environment is based on a Dell computer with Intel(R) Core(TM) i7-8700 CPU@3.20 GHz and 8GB RAM. We employ python 3.6.5 with XGBoost 0.9, Tensorflow 2.0, and Keras 2.2 to implement relevant models.

B. MEASURES OF EFFECTIVENESS

To evaluate the performance of the CEEMDAN-XGBoost model and benchmark models, the evaluation measures of the experiments include mean absolute error, root mean square error, mean absolute percentage error, and equality coefficient [57]. Taken together, the four measures evaluate an assessment to be made of accuracy and precision of prediction reference. These measures are defined as follows.

Mean absolute error (MAE):

$$\mathrm{MAE}=\frac{1}{S}\sum_{t=1}^{S}\left|\hat{y}_t-y_t\right| \tag{17}$$

Root mean square error (RMSE):

$$\mathrm{RMSE}=\sqrt{\frac{1}{S}\sum_{t=1}^{S}(\left|\hat{y}_t-y_t\right|)^2} \tag{18}$$

Mean absolute percentage error (MAPE):

$$\mathrm{MAPE}=\frac{1}{S}\sum_{t=1}^{S}\left|\frac{\hat{y}_t-y_t}{y_t}\right|\times 100\% \tag{19}$$

Equality coefficient (EC):

$$\mathrm{EC}=1-\frac{\sqrt{\sum_{t=1}^{S}\left(\hat{y}_t-y_t\right)^2}}{\sqrt{\sum_{t=1}^{S}\left(\hat{y}_t\right)^2}+\sqrt{\sum_{t=1}^{S}\left(y_t\right)^2}} \tag{20}$$

where $\hat{y}_t$ is the predicted traffic flow and y_t is the actual traffic flow. S is the number of the prediction time intervals.

Note that the MAE, RMSE, and MAPE represent the accuracy of the prediction models. EC reflects the prediction stability and fitting degree.

C. BENCHMARK MODELS

In order to demonstrate the superiority of the proposed model, ARIMA, MLP, BPNN, LSTM, Gated Recurrent Unit neural network (GRU), XGBoost, Wavelet-XGBoost, EMD-XGBoost, and EEMD-XGBoost are selected as benchmark models. Note that the decomposition-based models including the Wavelet-XGBoost, the EMD-XGBoost, and the EEMD-XGBoost are utilized to compare the performance of

different decomposition approaches including the Wavelet, the EMD, and EEMD, and the CEEMDAN approach. The relevant parameters and settings of the baseline models are illustrated as follows.

For the ARIMA, the optimal order p, d, q are determined by the best Akaike information criterion (AIC) value using the time series data in the two training datasets. For the traditional non-parametric such as the MLP, the BPNN, and the XGBoost model, the traffic flow of target lanes and their adjacent lanes at the same road section during previous ten-time intervals are utilized as the input variables of these models considering the strong relevance of the traffic flow between lanes shown in Figure.4. Besides, the MLP selects the following architecture: 30 units in the input layer, two hidden layers with 60 units in each hidden layers, and one unit in the output layer. The activation function is set as Relu function. The BPNN has an input layer with 30 units, a hidden layer with 60 units and an output layer with one unit. For the deep learning models including the LSTM and the GRU which can both capture nonlinear traffic dynamic effectively and automatically determine the optimal time lags, the number of the hidden layer is 100 and the largest time lags are chosen to be 10. The optimizers of two deep learning models are both set as Adam which uses an adaptive learning rate for stochastic optimization.

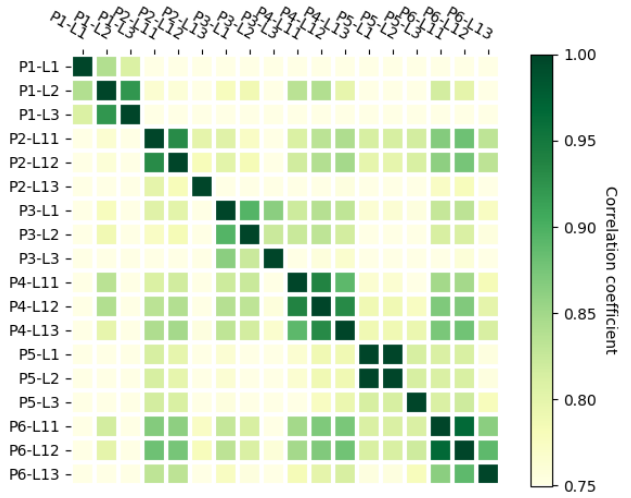


FIGURE 4. Correlation analysis of the traffic flow at different lanes.

For the XGBoost-based models including the XGBoost, the Wavelet-XGBoost, the EMD-XGBoost, the EEMD-XGBoost, and the CEEMDAN-XGBoost model, the learning rate and the maximum depth of a tree are selected as 0.1 and 6 respectively. The early stopping round is 20 and the objective function is to minimize the root mean square error between the predicted values and the ground-truth values. Note that the noise standard deviation of the EEMD and CEEMDAN are both set as 0.2. In addition, the numbers of realizations of the EEMD and CEEMDAN are chosen to be 500, and the maximum number of sifting iterations of the above two

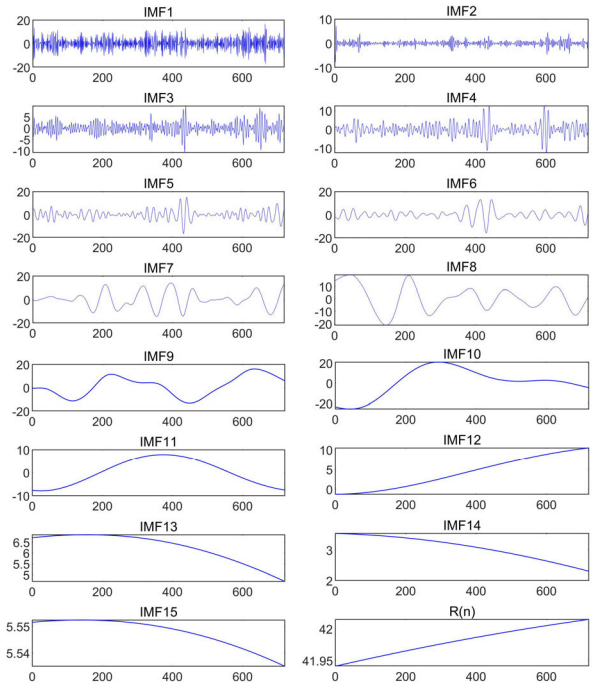


FIGURE 5. The decomposition result of the CEEMDAN method at lane section P1-L1 during a day (2014.1.6).

decomposition methods are 100. Meanwhile, the wavelet type of the Wavelet-XGBoost is set as db [43].

V. RESULTS AND DISCUSSION

A. THE RESULT OF THE CEEMDAN METHOD

Figure.5 reveals the decomposition results of the traffic flow during a day at the observation location P1-L1. It can be found that complicated original traffic data shown in Figure.3 is decomposed into 15 low-noise IMF components and one residual component. If these components in Figure.5 are synthesized, the wave of the original traffic flow in Figure.3 can still be obtained. If the prediction errors on each simple wave are lower, the prediction results achieved by integrating the results of simple waves will be more precise than the results forecasted based on the complicated original wave. Therefore, it is supposed that the decomposition-based models enable to decrease the unpredictability of lane-level traffic flow prediction through making predictions on the regular decomposed waves of the traffic flow.

B. COMPARISONS OF THE PREDICTION RESULTS OF DIFFERENT MODELS

In this section, Table 2 shows the overall performance of the different models in the task of making a single-step-ahead prediction by taking the results of all lanes as a whole. Among the single models, different models reveal pretty similar performance in the aspect of prediction accuracy and fitting degree. Meanwhile, it is interesting to find that the MAE, MAPE, and RMSE of the ARIMA model are slightly lower than those of the non-parametric methods (MLP, BPNN,

TABLE 2. Comparison of the overall performance of different models.

Methods	MAE	MAPE	RMSE	EC
ARIMA	4.21	21.34%	5.78	0.9242
MLP	4.27	23.06%	5.84	0.9234
BPNN	4.22	22.27%	5.78	0.9241
LSTM	4.24	21.30%	5.81	0.9239
GRU	4.27	21.45%	5.83	0.9231
XGBoost	4.26	22.77%	5.82	0.9236
Wavelet-XGBoost	2.50	14.66%	3.85	0.9495
EMD-XGBoost	2.56	14.37%	3.54	0.9537
EEMD-XGBoost	2.04	11.43%	2.86	0.9626
CEEMDAN-XGBoost	**1.79**	**9.88%**	**2.54**	**0.9668**

and XGBoost) and deep learning models (LSTM and GRU). Meanwhile, the models such as the MLP, the BPNN, and the XGBoost model which consider the spatio-temporal characteristics of the traffic flow don't show obvious advantages compared with the ARIMA, LSTM and GRU which only learn the temporal characteristics of the traffic flow.

However, compared with the single models using raw traffic flow data, the decomposition-based hybrid models demonstrate huge improvements on the MAEs, MAPEs, RMSEs, and ECs. The EMD-XGBoost model outperforms the XGBoost model with improvements 8.40% and 2.28 on MAPE and RMSE respectively.

In addition, the proposed CEEMDAN-XGBoost model illustrates the best performance among the hybrid models. The MAE, MAPE, and RMSE of the CEEMDAN-XGBoost model are 0.25, 1.55% and 0.32 lower than those of the EEMD-XGBoost model which reveals the second-best performance among the eight models. Meanwhile, the EC of the CEEMDAN-XGBoost model is 0.0042 more than that of the EEMD-XGBoost model. The experimental results also prove that the decomposition effectiveness of the CEEMDAN approach is more efficient than that of the Wavelet approach, the EMD approach and the EEMD approach. Hence the CEEMDAN approach can improve the accuracy and fitting degree of the XGBoost model more significantly.

To compare the prediction results of different models at different lanes, the lanes are divided into three types including the inside lanes, middle lanes, and outside lanes according to their physical characteristics. Table 3 shows the overall prediction results of different models at the above three types of lanes. It is demonstrated from the MAPEs and ECs that the accuracy and the fitting degree of different models at the inside lanes and middle lanes are superior to those of the models at the outside lanes. The possible reason for this phenomenon is that the frequent acceleration and deceleration of merging vehicles at the outside lanes affect the continuity and consistency of the traffic flow, which makes the traffic flow more unstable and difficult to forecast.

As shown in Table 3, the CEEMDAN-XGBoost model always has the lowest MAE, MAPE, and RMSE at different lanes. Even at the outside lanes, the EC of the CEEMDAN-XGBoost model can reach up to 0.9633, which is even higher

TABLE 3. Comparison of the overall performance of different models at different types of lanes.

Methods	Inside lane			
	MAE	MAPE	RMSE	EC
ARIMA	4.44	22.64%	6.32	0.9242
MLP	4.56	25.35%	6.45	0.9225
BPNN	4.48	23.93%	6.34	0.9237
LSTM	4.47	22.31%	6.33	0.9238
GRU	4.50	22.91%	6.37	0.9232
XGBoost	4.50	24.42%	6.35	0.9237
Wavelet-XGBoost	2.82	18.15%	4.38	0.9476
EMD-XGBoost	2.65	14.80%	3.78	0.9548
EEMD-XGBoost	2.11	11.89%	3.07	0.9632
CEEMDAN-XGBoost	**1.88**	**10.47%**	**2.79**	**0.9666**
Methods	Middle lane			
	MAE	MAPE	RMSE	EC
ARIMA	4.11	18.56%	5.49	0.9303
MLP	4.13	19.70%	5.46	0.9306
BPNN	4.10	19.16%	5.44	0.9308
LSTM	4.15	18.93%	5.52	0.9301
GRU	4.19	18.84%	5.56	0.9290
XGBoost	4.16	19.82%	5.53	0.9296
Wavelet-XGBoost	2.31	11.44%	3.50	0.9554
EMD-XGBoost	2.37	10.86%	3.23	0.9590
EEMD-XGBoost	1.93	9.24%	2.64	0.9664
CEEMDAN-XGBoost	**1.69**	**8.26%**	**2.36**	**0.9701**
Methods	Outside lane			
	MAE	MAPE	RMSE	EC
ARIMA	4.06	22.83%	5.50	0.9165
MLP	4.13	24.13%	5.55	0.9155
BPNN	4.09	23.71%	5.51	0.9162
LSTM	4.10	22.64%	5.53	0.9159
GRU	4.11	22.61%	5.54	0.9152
XGBoost	4.12	24.06%	5.54	0.9156
Wavelet-XGBoost	2.37	14.40%	3.61	0.9450
EMD-XGBoost	2.65	17.45%	3.59	0.9454
EEMD-XGBoost	2.09	13.14%	2.84	0.9568
CEEMDAN-XGBoost	**1.79**	**10.91%**	**2.44**	**0.9633**

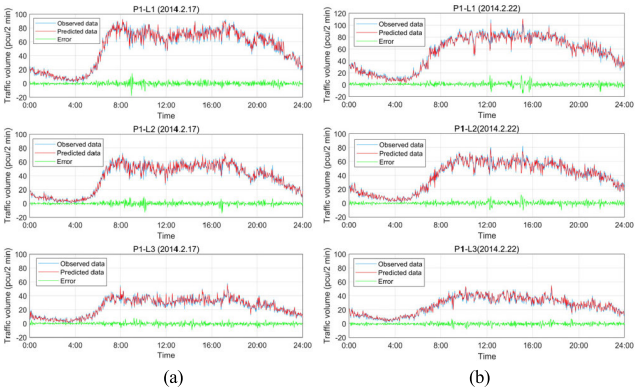


FIGURE 6. The prediction results of the lane-level traffic flow of the CEEMDAN-XGBoost model at P1. (a) On a workday; (b) On a weekend.

than that of the Wavelet-XGBoost and the EEMD-XGBoost model at the middle lanes. Therefore, the CEEMDAN-XGBoost model can effectively fit the complex nonlinear changes in traffic flow at different lanes and predict the traffic flow in the future accurately and stably.

Figure.6 presents the prediction results of the CEEMDAN-XGBoost model at three lanes on a workday (2014.2.17) and on a weekend (2014.2.22). In Figure.6, the

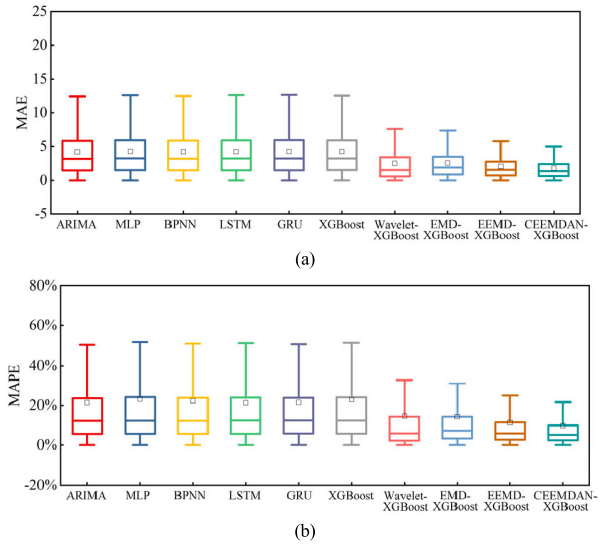


FIGURE 7. Comparison of the prediction errors for different methods (a) MAE; (b) MAPE.

CEEMDAN-XGBoost model is capable of capturing the tendency and volatility of the traffic flows of the different lane sections during the whole day. Even during the morning and evening peak periods on the workday when the traffic volume fluctuates greatly due to the complex traffic condition, the CEEMDAN-XGBoost model still has a dominant performance in fitting the sudden change of traffic volume and maintaining the prediction errors stable.

Figure.7 gives the overall prediction errors produced by these different methods in several lane sections. As shown in Figure.7, the proposed model outperforms other models in terms of the maximum, the minimum and the median of errors. Besides, it can be found that the CEEMDAN-XGBoost model has a smaller distance between $Q1$ and $Q3$ and the error distributions of the CEEMDAN-XGBoost model are more concentrated than those of other models. Hence, the MAE and MAPE of the proposed method shown in Figure.7 are lower than those of other methods, indicating that the CEEMDAN-XGBoost model is more precise and stable.

Figure.8 illustrates the correlation distribution between the actual values and the predicted values from six models including the ARIMA, the BPNN, the LSTM, the XGBoost, the EMD-XGBoost, and the CEEMDAN-XGBoost model in seven days. R^2 represents the correlation coefficient to evaluate the relevance between observed results and predicted results. From the observation in Figure.8, it can be found that the distributions of traffic flow sampling data are continuous from the low values to the high values. Moreover, it can be concluded that the CEEMDAN-XGBoost model produces better prediction results with higher R^2 compared with other models. In addition, the EMD-XGBoost achieves higher R^2 than four single models and it indicates the common superiority of the decomposition-based methods.

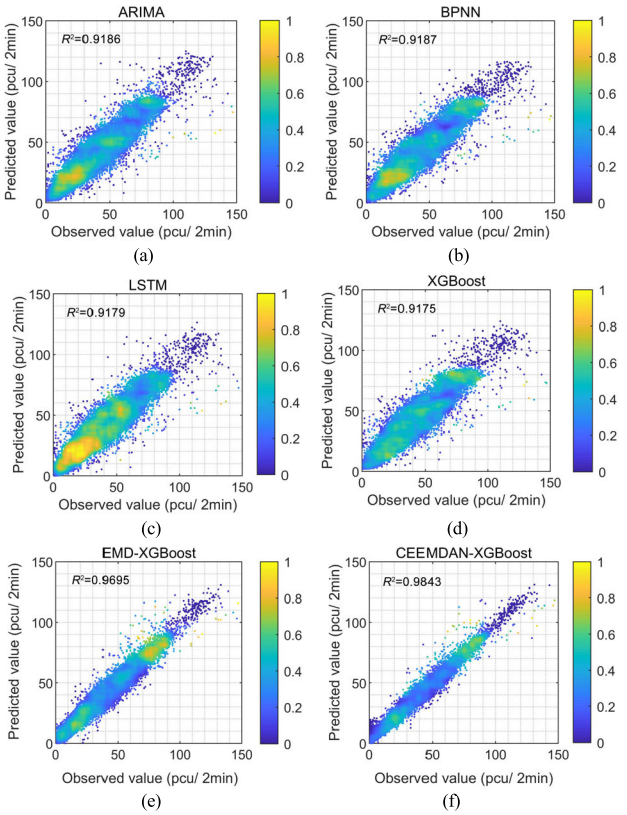


FIGURE 8. Correlation distribution of the predicted results of the different models (a) ARIMA; (b)BPNN; (c) LSTM; (d) XGBoost; (e) EMD-XGBoost; (f) CEEMDAN-XGBoost.

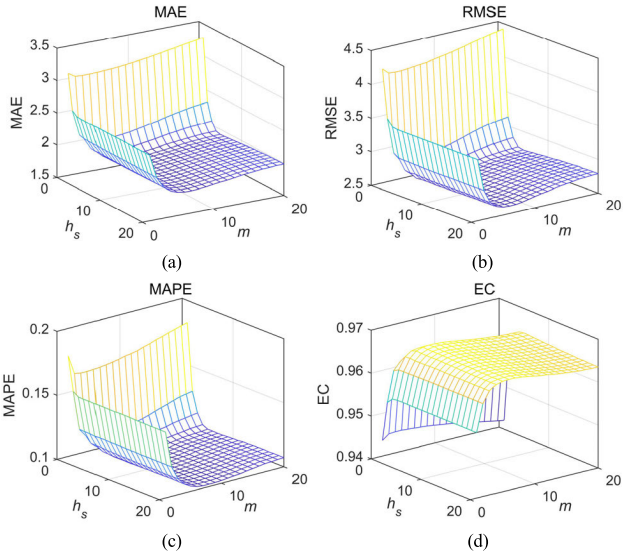


FIGURE 9. Prediction performance of the CEEMDAN-XGBoost model with different parameters. (a) MAE; (b) RMSE; (c) MAPE; (d) EC.

C. PARAMETER ANALYSIS OF THE CEEMDAN-XGBOOST MODEL

In the framework of the proposed model, the input time step h_s and the maximum depth of a tree m in the XGBoost algorithm are the two critical parameters which may affect the prediction performance of the CEEMDAN-XGBoost model. Therefore, as shown in Figure.9, we examine the

CEEMDAN-XGBoost models with different settings of the two parameters. The input time step h_s is tuned from 1 to 20 with a step of 1, which means the input time horizon increase from 2 min to 40 min. Meanwhile, the maximum depth of a tree m is tuned from 1 to 20 with a step of 1.

Figure.9 indicates that when the input time step is fixed, the MAEs, RMSEs, and MAPEs firstly drop and then rise a little with the increase of the maximum depth of a tree. On the contrary, the ECs firstly rise and then drop a little with the increase of the maximum depth of a tree. In addition, when the maximum depth of a tree is fixed, the MAE, RMSE, and MAPE firstly decrease and then stable with the increase of the input time step. It can be learned from Figure.9 that the recommended values of the input time step h_s and the maximum depth of a tree m both range from 5 to 10 considering the time consumption and prediction performance of the model.

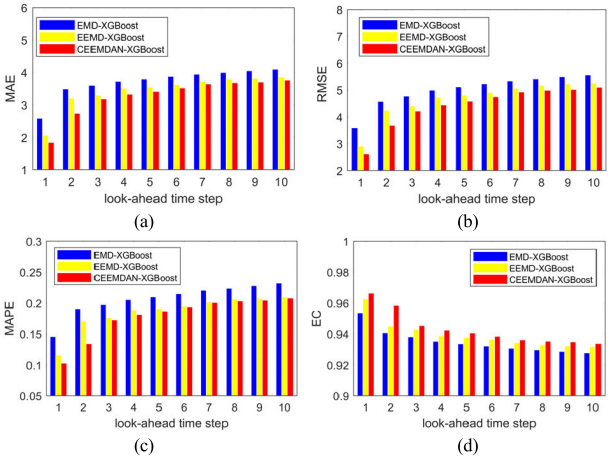


FIGURE 10. Prediction performance of the XGBoost-based models in longer prediction horizon. (a) MAE; (b) RMSE; (c) MAPE; (d) EC.

D. COMPARISON OF THE XGBOOST-BASED MODELS IN THE MULTI-STEP-AHEAD PREDICTION

Forecasting traffic flow over several time intervals in the future allows a wider range of applications to take advantage of predictions. Figure.10 compares the prediction results of the XGBoost-based models with the look-ahead time step increasing from 1 to 10, which corresponds to the prediction horizon ranging from 2 min to 20 min. In general, the accuracy and the fitting degree of the CEEMDAN-XGBoost model deteriorate slightly with the increase of the look-ahead time step, and the CEEMDAN-XGBoost model shows better performance than the EMD-XGBoost and EEMD-XGBoost model in terms of the MAE, RMSE, MAPE, and EC. It is worth noting that when the prediction horizon reaches up to 20 min, the prediction performance of the proposed model is still better than that of the single models including the ARIMA, BPNN, MLP, XGBoost, LSTM, and GRU with the prediction horizon set as 2 min shown in TABLE 1. Hence, the CEEMDAN-XGBoost model can deal with multi-step-ahead traffic prediction efficiently.

VI. CONCLUSION AND FUTURE WORK

Reliable lane-level short-term traffic prediction is of critical importance for both CAVs and HDVs in the CAVH system. This paper presents a novel hybrid model for short-term traffic flow prediction by fusing complete ensemble empirical mode decomposition with adaptive noise and extreme gradient boosting algorithm. The CEEMDAN method is utilized to decompose the traffic flow into multiple highly predictable IMF components including one residual component. The decomposed components are sent to the XGBoost algorithm for training and forecasting separately. At last, the predicted values of all decomposed components are accumulated to obtain the predicted traffic flow of the proposed model. To validate the effectiveness of the proposed CEEMDAN-XGBoost model, the real-world traffic flow data from the 3rd Ring Road of Beijing were collected by RTMSs to conduct experiments with the data of the first two weeks for training and the data of the rest week for testing. In addition, the ARIMA, MLP, BPNN, LSTM, GRU, XGBoost, Wavelet-XGBoost, EMD-XGBoost, and EEMD-XGBoost model were employed as the benchmark models. The performance of the proposed method was evaluated in terms of four measurement criteria: MAE, RMSE, MAPE, and EC.

Several useful findings and recommendations can be generated in this study. (1) Experimental results indicate that the CEEMDAN-XGBoost model can efficiently learn and capture the traffic patterns under different traffic condition at different types of lanes. (2) Comparisons with benchmark models reveal that the proposed model achieves superior prediction accuracy and stability over the ARIMA, MLP, BPNN, LSTM, GRU, XGBoost, Wavelet-XGBoost, EMD-XGBoost, and EEMD-XGBoost model in the task of making short-term lane-level traffic flow prediction. (3) The input time step and the maximum depth of a tree both affect the prediction performance of the CEEMDAN-XGBoost model. The proposed model obtains its satisfactory prediction performance with the two parameters both ranging from 5 to 10. (4) The CEEMDAN-XGBoost model outperforms the other XGBoost-based models such as the Wavelet-XGBoost, EMD-XGBoost and EEMD-XGBoost model in the task of forecasting both single-step-ahead and multi-step-ahead lane-level traffic flow.

The future work will be conducted by considering both spatial and temporal information into the XGBoost models. The periodic characteristics of the original traffic flow can be mined to improve the performance of the proposed model. In addition, the CEEMDAN-XGBoost model provides another unique perspective on solving the complex nonlinear issues such as the vehicle trajectory prediction and the travel time estimation.

REFERENCES

[1] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-term traffic forecasting: Where we are and where we're going," *Transp. Res. C, Emerg. Technol.*, vol. 43, no. 1, pp. 3–19, Jun. 2014.

[2] M. G. Karlaftis and E. I. Vlahogianni, "Statistical methods versus neural networks in transportation research: Differences, similarities and some insights," *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 3, pp. 387–399, Jun. 2011.

[3] B. Ran, "Connected automated vehicle highway systems and methods," U.S. Patent Appl. 10 380 886 B2, Aug. 8, 2019.

[4] Y. Gu, W. Lu, X. Xu, L. Qin, Z. Shao, and H. Zhang, "An improved Bayesian combination model for short-term traffic prediction with deep learning," *IEEE Trans. Intell. Transp. Syst.*, to be published.

[5] Z. Yao, R. Hu, Y. Wang, Y. Jiang, B. Ran, and Y. Chen, "Stability analysis and the fundamental diagram for mixed connected automated and human-driven vehicles," *Phys. A, Stat. Mech. Appl.*, vol. 533, Nov. 2019, Art. no. 121931.

[6] T. Song, N. Capurso, X. Cheng, J. Yu, B. Chen, and W. Zhao, "Enhancing GPS with lane-level navigation to facilitate highway driving," *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 4579–4591, Jun. 2017.

[7] D. Tian, G. Wu, P. Hao, K. Boriboonsomsin, and M. J. Barth, "Connected vehicle-based lane selection assistance application," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 7, pp. 2630–2643, Jul. 2019.

[8] C. F. Daganzo, "A behavioral theory of multi-lane traffic flow. Part I: Long homogeneous freeway sections," *Transp. Res. B, Methodol.*, vol. 36, no. 2, pp. 131–158, Feb. 2002.

[9] C. F. Daganzo, "A behavioral theory of multi-lane traffic flow. Part II: Merges and the onset of congestion," *Transp. Res. B, Methodol.*, vol. 36, no. 2, pp. 159–169, Feb. 2002.

[10] P. G. Michalopoulos, D. E. Beskos, and Y. Yamauchi, "Multilane traffic flow dynamics: Some macroscopic considerations," *Transp. Res. B, Methodol.*, vol. 18, nos. 4–5, pp. 377–395, Aug. 1984.

[11] Y. Xuan, C. F. Daganzo, and M. J. Cassidy, "Increasing the capacity of signalized intersections with separate left turn phases," *Transp. Res. B, Methodol.*, vol. 45, no. 5, pp. 769–781, Jun. 2011.

[12] K. Yuan, V. L. Knoop, L. Leclercq, and S. P. Hoogendoorn, "Capacity drop: A comparison between stop-and-go wave and standing queue at lane-drop bottleneck," *Transportmetrica B, Transp. Dyn.*, vol. 5, no. 2, pp. 145–158, Nov. 2016.

[13] Y. Gu, W. Lu, L. Qin, M. Li, and Z. Shao, "Short-term prediction of lane-level traffic speeds: A fusion deep learning model," *Transp. Res. C, Emerg. Technol.*, vol. 106, pp. 1–16, Sep. 2019.

[14] Y. Wu, H. Tan, L. Qin, B. Ran, and Z. Jiang, "A hybrid deep learning based traffic flow prediction method and its understanding," *Transp. Res. C, Emerg. Technol.*, vol. 90, pp. 166–180, May 2018.

[15] D. Zang, Y. Fang, Z. Wei, K. Tang, and J. Cheng, "Traffic flow data prediction using residual deconvolution based deep generative network," *IEEE Access*, vol. 7, pp. 71311–71322, 2019.

[16] Y. Xie, Y. Zhang, and Z. Ye, "Short-term traffic volume forecasting using Kalman filter with discrete wavelet decomposition," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 22, no. 5, pp. 326–334, Jul. 2007.

[17] K. Y. Chan, T. S. Dillon, J. Singh, and E. Chang, "Neural-network-based models for short-term traffic flow forecasting using a hybrid exponential smoothing and Levenberg–Marquardt algorithm," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 2, pp. 644–654, Jun. 2012.

[18] M. Levin and Y.-D. Tsao, "On forecasting freeway occupancies and volumes," *Transp. Res. Rec.*, vol. 773, no. 1, pp. 47–49, 1980.

[19] B. Ghosh, B. Basu, and M. O'Mahony, "Multivariate short-term traffic flow forecasting using time-series analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 2, pp. 246–254, Jun. 2009.

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

[21] M. Van Der Voort, M. Dougherty, and S. Watson, "Combining Kohonen maps with ARIMA time series models to forecast traffic flow," *Transp. Res. C, Emerg. Technol.*, vol. 4, no. 5, pp. 307–318, Oct. 1996.

[22] S. Lee and D. B. Fambro, "Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1678, no. 1, pp. 179–188, Jan. 1999.

[23] B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results," *J. Transp. Eng.*, vol. 129, no. 6, pp. 664–672, Nov. 2003.

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

[25] B. L. Smith, B. M. Williams, and R. Keith Oswald, "Comparison of parametric and nonparametric models for traffic flow forecasting," *Transp. Res. C, Emerg. Technol.*, vol. 10, no. 4, pp. 303–321, Aug. 2002.

[26] H. Jiang, Y. Zou, S. Zhang, J. Tang, and Y. Wang, "Short-term speed prediction using remote microwave sensor data: Machine learning versus statistical model," *Math. Problems Eng.*, vol. 2016, pp. 1–13, Feb. 2016.

[27] W.-C. Hong, Y. Dong, F. Zheng, and C.-Y. Lai, "Forecasting urban traffic flow by SVR with continuous ACO," *Appl. Math. Model.*, vol. 35, no. 3, pp. 1282–1291, Mar. 2011.

[28] O. A. Arqub, M. Al-Smadi, S. Momani, and T. Hayat, "Application of reproducing kernel algorithm for solving second-order, two-point fuzzy boundary value problems," *Soft Comput.*, vol. 21, no. 23, pp. 7191–7206, Jul. 2016.

[29] H. Chang, Y. Lee, B. Yoon, and S. Baek, "Dynamic near-term traffic flow prediction: System-oriented approach based on past experiences," *IET Intell. Transp. Syst.*, vol. 6, no. 3, pp. 292–305, 2012.

[30] Z. Su, Q. Liu, J. Lu, Y. Cai, H. Jiang, and L. Wahab, "Short-time traffic state forecasting using adaptive neighborhood selection based on expansion strategy," *IEEE Access*, vol. 6, pp. 48210–48223, 2018.

[31] D. Xia, H. Li, B. Wang, Y. Li, and Z. Zhang, "A map reduce-based nearest neighbor approach for Big-Data-Driven traffic flow prediction," *IEEE Access*, vol. 4, pp. 2920–2934, 2016.

[32] D. Park and L. R. Rilett, "Forecasting freeway link travel times with a multilayer feedforward neural network," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 14, no. 5, pp. 357–367, Sep. 1999.

[33] J. W. C. van Lint, S. P. Hoogendoorn, and H. J. van Zuylen, "Accurate freeway travel time prediction with state-space neural networks under missing data," *Transp. Res. C, Emerg. Technol.*, vol. 13, nos. 5–6, pp. 347–369, Oct. 2005.

[34] X. Zeng and Y. Zhang, "Development of recurrent neural network considering temporal-spatial input dynamics for freeway travel time modeling," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 28, no. 5, pp. 359–371, Jan. 2013.

[35] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Spatio-temporal short-term urban traffic volume forecasting using genetically optimized modular networks," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 22, no. 5, pp. 317–325, Jul. 2007.

[36] J. Tang, F. Liu, Y. Zou, W. Zhang, and Y. Wang, "An improved fuzzy neural network for traffic speed prediction considering periodic characteristic," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 9, pp. 2340–2350, Sep. 2017.

[37] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," *Transp. Res. C, Emerg. Technol.*, vol. 54, pp. 187–197, May 2015.

[38] Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen, and J. Liu, "LSTM network: A deep learning approach for short-term traffic forecast," *IET Intell. Transp. Syst.*, vol. 11, no. 2, pp. 68–75, Mar. 2017.

[39] H.-F. Yang, T. S. Dillon, and Y.-P.-P. Chen, "Optimized structure of the traffic flow forecasting model with a deep learning approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2371–2381, Oct. 2017.

[40] J. Wang, W. Deng, and Y. Guo, "New Bayesian combination method for short-term traffic flow forecasting," *Transp. Res. C, Emerg. Technol.*, vol. 43, pp. 79–94, Jun. 2014.

[41] Y. Zhang, Y. Zhang, and A. Haghani, "A hybrid short-term traffic flow forecasting method based on spectral analysis and statistical volatility model," *Transp. Res. C, Emerg. Technol.*, vol. 43, pp. 65–78, Jun. 2014.

[42] F. Guo, J. W. Polak, and R. Krishnan, "Predictor fusion for short-term traffic forecasting," *Transp. Res. C, Emerg. Technol.*, vol. 92, pp. 90–100, Jul. 2018.

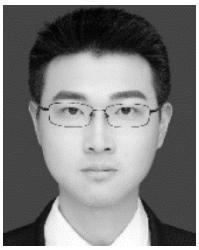
[43] J. Tang, X. Chen, Z. Hu, F. Zong, C. Han, and L. Li, "Traffic flow prediction based on combination of support vector machine and data denoising schemes," *Phys. A, Stat. Mech. Appl.*, vol. 534, Nov. 2019, Art. no. 120642.

[44] E. I. Vlahogianni, "Optimization of traffic forecasting: Intelligent surrogate modeling," *Transp. Res. C, Emerg. Technol.*, vol. 55, pp. 14–23, Jun. 2015.

[45] L. Li, L. Qin, X. Qu, J. Zhang, Y. Wang, and B. Ran, "Day-ahead traffic flow forecasting based on a deep belief network optimized by the multi-objective particle swarm algorithm," *Knowl.-Based Syst.*, vol. 172, no. 15, pp. 1–14, May 2019.

[46] H. Wang, L. Liu, S. Dong, Z. Qian, and H. Wei, "A novel work zone short-term vehicle-type specific traffic speed prediction model through the hybrid EMD–ARIMA framework," *Transportmetrica B, Transp. Dyn.*, vol. 4, no. 3, pp. 159–186, Jul. 2015.

[47] L. Li, X. Qu, J. Zhang, H. Li, and B. Ran, "Travel time prediction for highway network based on the ensemble empirical mode decomposition and random vector functional link network," *Appl. Soft Comput.*, vol. 73, pp. 921–932, Dec. 2018.
[48] Y. Wei and M.-C. Chen, "Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks," *Transp. Res. C, Emerg. Technol.*, vol. 21, no. 1, pp. 148–162, Apr. 2012.
[49] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, "A complete ensemble empirical mode decomposition with adaptive noise," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 4144–4147.
[50] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: A noise-assisted data analysis method," *Adv. Adapt. Data Anal.*, vol. 1, no. 1, pp. 1–41, Nov. 2011.
[51] P. Flandrin, G. Rilling, and P. Goncalves, "Empirical mode decomposition as a filter bank," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 112–114, Feb. 2004.
[52] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. London. A, Math., Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, Mar. 1998.
[53] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2016, pp. 785–794.
[54] J. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1190–1232, 2000.
[55] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *Ann. Statist.*, vol. 28, no. 2, pp. 337–374, 2000.
[56] F. Zhan, X. Wan, Y. Cheng, and B. Ran, "Methods for multi-type sensor allocations along a freeway corridor," *IEEE Intell. Transp. Syst. Mag.*, vol. 10, no. 2, pp. 134–149, Apr. 2018.
[57] J. Wang and Q. Shi, "Short-term traffic speed forecasting hybrid model based on Chaos–Wavelet analysis-support vector machine theory," *Transp. Res. C, Emerg. Technol.*, vol. 27, no. 2, pp. 219–232, Feb. 2013.



WENQI LU received the B.E. degree in transportation engineering from Hohai University, Nanjing, China, in 2016, and the M.S. degree in transportation planning and management from Beijing Jiaotong University, Beijing, China, in 2019. He is currently pursuing the Ph.D. degree with the School of Transportation, Southeast University, Nanjing.

His current research interests include connected and automated vehicles, traffic theory, and intelligent transportation systems.

YIKANG RUI received the Ph.D. degree in geographic information science from the KTH Royal Institute of Technology.

He was a Postdoctoral Fellow with Nanjing University, from 2014 to 2016. He is currently an Associate Research Fellow with the School of Transportation, Southeast University, Nanjing, China. His current research interests include spatio-temporal data model, traffic network analysis, and vehicle-road collaborative decision planning.

ZIWEI YI received the B.E. degree in transportation engineering from the Harbin Institute of Technology, Harbin, China, in 2017. She is currently pursuing the Ph.D. degree with the School of Transportation, Southeast University, Nanjing, China. Her research interests include traffic flow theory and intelligent vehicles.

BIN RAN received the Ph.D. degree from the University of Illinois at Chicago, USA, in 1993. He is currently a Professor with the Department of Civil and Environmental Engineering, University of Wisconsin–Madison, WI, USA, and the Director of the Research Center for Internet of Mobility, Southeast University, Nanjing, China. He is one of the co-founders of the Chinese Overseas Transportation Association, and he was the first Chairman. He has authored or coauthored over 90 articles in international journals, including *Transportation Science*, *Transportation Research Part B*, and IEEE Access.

YUANLI GU received the Ph.D. degree in transportation planning and management from Beijing Jiaotong University, Beijing, China, in 2010. He was a Visiting Scholar with the University of Wisconsin–Madison, Madison, WI, USA, from 2016 to 2017. He is currently an Associate Professor with the MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology, Beijing Jiaotong University. His current research interests include traffic planning, traffic management and control, and intelligent transportation systems.