IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ARTIFICIAL INTELLIGENCE (AI)-EMPOWERED INTELLIGENT TRANSPORTATION SYSTEMS

Received March 16, 2020, accepted April 16, 2020, date of publication May 6, 2020, date of current version May 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2992657

A Period-Specific Combined Traffic Flow Prediction Based on Travel Speed Clustering

BIN FENG[1], JIANMIN XU[1], YONGJIE LIN[1], (Member, IEEE), AND PENGHAO LI[2]
[1]School of Civil Engineering and Transportation, South China University of Technology, Guangzhou 510641, China
[2]Traffic Police Detachment, Zhongshan 528403, China

Corresponding author: Yongjie Lin (linyjscut@scut.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1601600, in part by the National Natural Science Foundation of China under Grant 61903145, and in part by the Natural Science Foundation of Guangdong Province under Grant 2018A030310395.

ABSTRACT Short-term traffic flow forecasting has always been an interesting research at the fields of Intelligent Transportation Systems. This paper presents a time-based combined traffic flow prediction model based on field data collected by loop detectors at signalized intersections, which are used to signal optimization, route choice, traffic monitoring, etc. Firstly, the traffic flow and corresponding travel speed by hour is processed for error elimination and correlation analysis. Secondly, time of day is divided into three groups (peak, flat-peak and low-peak period) in terms of hourly travel speed clustering such as to separately develop prediction formula for each period with avoiding the overfitting of a single 24-hour model. And then, a combined prediction model based on time partition is proposed for 24-hour traffic flow forecasting, which adopts grey theory model for flat-peak and low-peak periods and back-propagation artificial neural network for peak hours, respectively. Finally, in tests that used field data from Xingzhong Rd, Zhongshan, China, the developed combined method based on speed clustering shows promise in reducing mean absolute error, mean absolute percentage error and mean squared error. Further exploration with excessive experiments for comparison analysis exhibits that the period-specific combined model conducts a more accurate and reliable prediction than the individual model and existing combined ones with the same structure for 24-hour.

INDEX TERMS Traffic flow prediction, cluster analysis, gray theory model, backpropagation artificial neural network.

I. INTRODUCTION

Urbanization development has been causing serious traffic congestion in numerous metropolitan and large cities around the world. Thus, it's necessary and inevitable to conduct urban road infrastructure construction and advanced traffic management for meeting travel demand [1]. Most effective strategies for traffic congestion mitigation always depend on the accurate and timely traffic prediction, such as traffic flow for traffic organization and signal timing optimization, travel time or speed for vehicle routing guidance.

Since early 1980s, short-term traffic prediction technique has become one of the most important components of Intelligent Transportation System (ITS), and the prediction time window ranges from a few minutes to a few hours into the future based on road geometry, traffic information, and control strategies, etc. [2]. In review of literatures over the past few decades, the forecasting model can be roughly classified into two categories: single models and combined ones.

The former is usually dedicated on one certain kind of formula by considering current and past traffic information. The existing literature can be divided into two subgroups. One category is parametric models, which can be described by using a finite number of parameters, such as exponential smoothing model [3], historical average algorithm [4], Autoregressive Integrated Moving Average (ARIMA) [5], Kalman Filtering (KF) [6], [7], and Grey theory model (GM) [8]. Among, GM is suitable to predict the system of having poor information and uncertainty, such as traffic flow [9]. Okutani and Stephanedes [10] began to employ Kalman filter to forecast traffic flow on the road network in Nagoya, Japan. Moreover, Sun *et al.* [11] developed a linear regression model to forecast flow on US-290 freeway in Houston, USA, who found that it outperformed the k-nearest neighbor method and kernel smoothing method. The other category, non-parametric models, assumes the structure of traffic

The associate editor coordinating the review of this manuscript and approving it for publication was Edith C.-H. Ngai.

parameters is not fixed and mainly follows statistical regularity depending on the abundant field data, such as Support Vector Machine (SVM) [1], artificial neural networks [12]–[14], non-parametric regression [15], Gaussian maximum likelihood [16]. Among, Artificial Neural Network (ANN) is one of the most widely used methods because it can capture traffic fluctuation [17]. Liang and Wei [18] modeled traffic flow on freeways based on simple recurrent networks, also namely Elman Network. Tan *et al.* [19] proved the k-Nearest Neighbor (k-NN) model can outperform ARIMA and the exponential smoothing model based on the field data of Guangzhou, China. Later, Lv *et al.* [20] first applied Stacked Auto-Encoder (SAE) to short-term traffic flow prediction and trained the model by Greedy Layer-Wise algorithm. Recently, Yu *et al.* [21] analyzed the temporal-spatial characteristics of traffic flow, and then employed Convolutional Neural Network (CNN) to predict short-term flow based on location partition. Subsequently, Xu *et al.* [22] developed an artificial fish swarm algorithm to optimize support vector machine regression for flow forecasting. With the development of computer techniques, machine learning has also been used for flow forecasting on the basis of huge historical dataset. For example, Dai *et al.* [23] proposed a many-to-many deep learning for traffic prediction, namely DeepTrend 2.0, which regards multi-sensor information as input and simultaneously generates predicted results for all sensors. Meanwhile, Li *et al.* [24] proposed a deep feature leaning approach in the following multiple steps by using supervised learning techniques. Zhao *et al.* [25] predicted traffic flow on four road segments in Beijing by using LSTM, and found it outperformed ARIMA and RNN (Recurrent Neural Network). Polson and Sokolov [26] pointed out that deep learning architectures are able to capture the nonlinear spatial-temporal effects resulting from the transitions between free flow, breakdown, recovery, and congestion in traffic flow.

Different from the previous single forecasting techniques, most combined models could yield much more benefits than the same kind individual ones due to utilizing two or more methods' advantages. For example, some researchers preferred to choosing ANN as the underlying model of integrating other methods, such as GM [27], [28], ARIMA [29], k-NN [30], Support Vector Regression (SVR) [31], clustering algorithm [32], and simple statistical approach [33]. Also, Feng *et al.* [34] combined wavelet function and Extreme Learning Machine (ELM) to propose a short-term prediction which outperformed ANN based on the field data of Canadian highway. Recently, Wu *et al.* [35] proposed a combined deep learning of CNN-RNN by considering the weekly/daily periodicity and spatial-temporal characteristics of traffic flow. According to one decomposed periodic sequence and two-part random ones for traffic flow time series, Zheng *et al.* [36] proposed a hybrid prediction with Back Propagation (BP)-based ANN, ε-SVR and LSTM models. Chang and Tsai [37] reported a composite method where incorporating a generalized auto-regressive conditional heteroscedasticity into GM tuned by adaptive support vector regression.

Although many worldwide researchers have reported a large variety of methods on traffic flow prediction, the short-term traffic flow forecasting along urban signalized corridors is still challenging tasks because traffic flows have many uncertainty (e.g. time-varying, highly oscillated, nonlinear and non-stationary). In particularly, most studies are dedicated on one single model for 24-hour forecasting by quantifying the relationship between the predictor and dependent input variables, which might result in model overfitting due to the high fluctuations of traffic flow over hours. Therefore, this study contributes to proposing a new period-based combined scheme of GM and BP based on the field historical datasets. The main contents of this paper can be divided into two parts: (i) The k-means clustering method is employed to divide 24 hours of one day into multiple time periods based on the travel speed time series in hour; and (ii) a combined method of GM and BP, namely GM-BP, is developed to forecast the hourly traffic flow for each time period, which can capture the fluctuation and overfitting prevention.

The remaining of the article is organized as follows: Section 2 describes the field dataset and data processing. In Section 3, the detailed prediction method is developed discussed. The case study demonstrates model performance with the field data from the city of Zhongshan in Section 4. The last section presents the conclusions.

II. DATA COLLECTION AND ANALYSIS

As known, it's significant to investigate traffic flow prediction based on actual traffic data. However, it's time-consuming and expensive for local governments or traffic engineers to collect the large-scale traffic data in practice. In the past decades, local governments in many Chinese cities have installed many infrastructures and developed application systems based on the idea of ITS.

A. DATA SOURCE

As one of the earliest pilot cities of ITS in China, the city of Zhongshan in Guangdong Province has ability to automatically collect the city-level traffic flow at signalized intersections. Therefore, this study collected hourly traffic flow and link travel speed belonging to ITS with Internet Plus from the department of Zhongshan Traffic Police Detachment.

In details, the tested site is located on Xingzhong Rd with two-way six motorized lanes, which is the busiest and most congested south-north corridors in Zhongshan downtown area. There are many government agencies, commercial buildings, and activity centers along Xingzhong Rd. The dataset with time interval of one hour was recorded from February 27 to March 26, 2017, and the total sample size is 672. Among, it included southbound traffic flow collected by loop detectors installed several meters before the southbound stop-line at the signalized intersection between Xingzhong Rd and Songyuan Rd, and link travel speed from Sunwen East Road to Songyuan Road along Xinzhong Road in Figure 1.

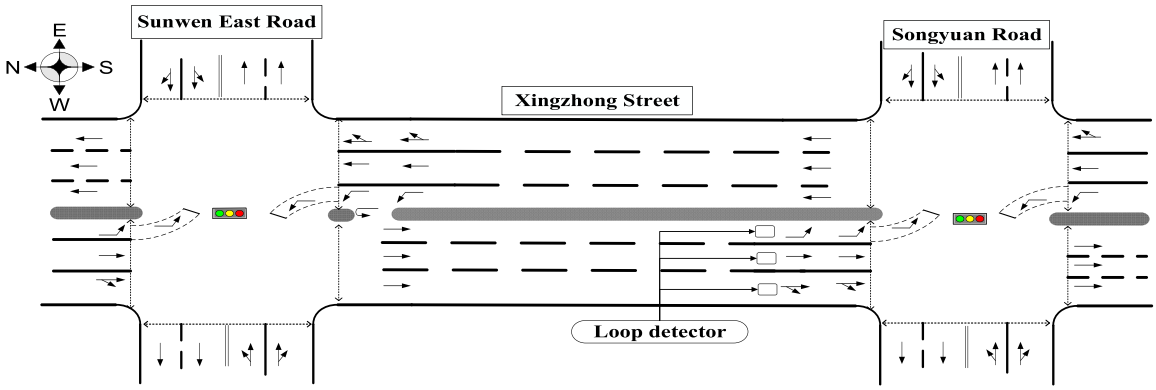


FIGURE 1. The layout of pilot intersections in the city of Zhongshan, China.

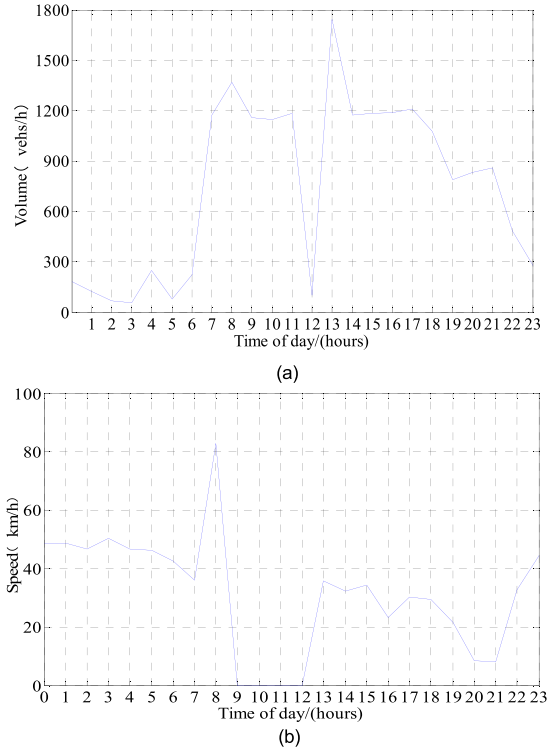


FIGURE 2. Abnormal data sample. (a) singular volume. (b) zero travel speed.

B. ABNORMAL DATA IDENTIFICATION

Based on the basic data analysis, one can find out that there are some abnormal data of raw traffic flow and average travel speed as depicted in Figure 2, which could be caused by the detector failure due to power off, communication interrupt, etc. In details, traffic volumes suddenly dropped from about 1200 vehs/h at 11:00 to 100 vehs/h at 12:00, and dramatically increased up to 1750 vehs/h at 13:00 on March 2. Similarly, average travel speed also reached to 82.8 km/h at 8:00 on February 11, and then dropped to 0 km/h in the next four hours. Thus, it's necessary to identify and eliminate these abnormal data before prediction.

Generally, the Wright criterion (i.e. 3σ criterion) [38] is a very effective method for discriminating outliers in the case of a normal distribution. This study proposed a data processing procedure based on this criterion. Firstly, let's define the residual between the hourly traffic volume and average one detected by loop detectors by:

$$\Delta q(i) = q(i) - \overline{q} \tag{1}$$

where, $q(i)$ represents the detected traffic volume at the ith hour; $\overline{q}$ is the mean of total sample data. If the absolute residual for the ith sample is greater than the triple standard deviation of the absolute residual, it will be marked as abnormal data which need be calibrated by other methods. This method is also applied for link travel speed in this paper.

C. DATA CORRELATION ANALYSIS

As known, there are many relevant variables of traffic flow prediction in the literature, such as historical flow [39], [40], travel speed [41], [42], traffic state [43], congestion levels [44], and occupancy [45]. Moreover, it's greatly expensive and difficult to collect traffic signal timing plan because sometimes it is adaptive or actuated based on control logic in practice. Therefore, this article contributes to developing a feasible prediction method based on the available data in Zhongshan. In order to determine the prediction formula and inputs, we firstly analyzed the triple-week dataset with traffic flow and travel speed in hour, and illustrated some key findings via one-week data from March 20 to March 26 as shown in Figure 3. And, Figure 4 shows link travel speed estimated by floating car data, which is from Sunwen East Road to Songyuan Road along Xinzhong Road in Figure 1.

During the analyzed time windows, traffic system is quite stable without special incidents along the targeted corridor, such as holiday, major events, and school opening or closing. Compared with Figures 3 and 4, traffic speed time series was quite stable and high from 22:00 to the next 6:00, namely late-night off-peak hours. And then, it was dropped and traffic congestion happened from 7:00 to 8:00, and from 17:00 to 18:00, namely morning and evening peak hours, respectively;

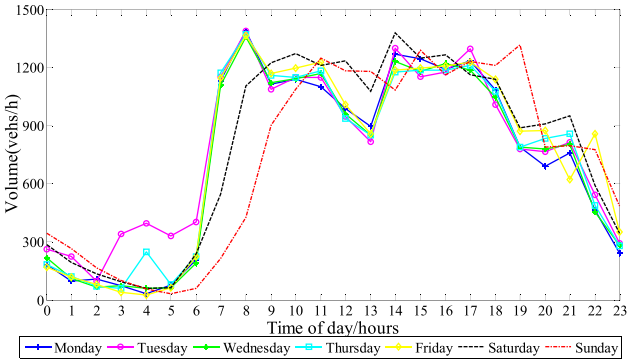


FIGURE 3. One-week southbound traffic flow distribution at the intersection between Xingzhong Rd and Songyuan Rd.

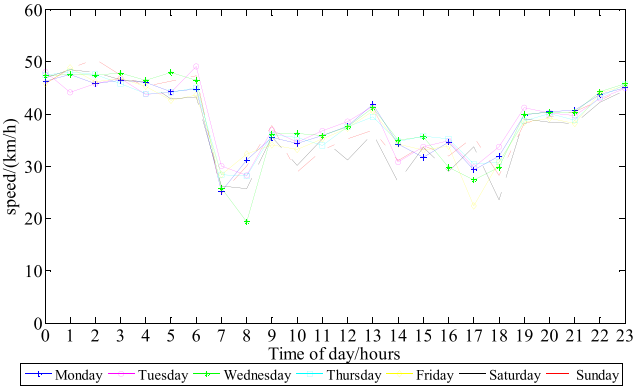


FIGURE 4. One-week link travel speed distribution from Sunwen East Road to Songyuan Road along Xinzhong Road.

and the remaining are regarded as transition process with the high fluctuation of volume and speed. In addition, there is a slight decline in speed at noon on Saturday and Sunday.

In order to further explore the characteristics of hourly traffic flow, a correlation analysis is conducted via Pearson coefficient by using the derived data from February 27 to March 26, 2017. Overall, on the same day, the correlation coefficients are getting more and more smaller with the increase of time difference in the last column of Table 1. The coefficients between adjacent two hours are greater than 0.8 over different time of day, and the coefficients of the last three intervals also exceed 0.5. The results showed that the current interval volume has a significant correlation with the past three ones, which should be considered into model development.

Meanwhile, one can also observe that the coefficients during the past one-week is much higher than 0.85 with regardless of the day of week in Table 2. Different from other roads in Zhongshan, the correlation between weekdays and weekends is still higher because many activity centers located on the either side of the arterial attract many local residents and Xingzhong Rd is also the most important south-north arterials for interzone travelers. Therefore, it's possible to use historical time series to estimate missing or abnormal data. Thus, this study presented that the abnormal or missing travel speed and flow would be set to the average value of the same time interval on the same day of the past three weeks.

III. MODEL DEVELOPMENT

A. MODEL STRUCTURE

As known, traffic flow cannot directly reflect traffic condition unless it combines with other parameters, such as the number of lanes, saturation flow rate, and signal timing. However, travel speed is a popular variable to effectively represent traffic congestion. Thus, this study presents a speed-cluster method to identify traffic congestion and decompose 24 hours of one day into multiple periods, and developed the specific flow prediction algorithm for each period, namely period-specific prediction. The scheme of the entire prediction logic is developed as follows:

1) DATA PROCESSING

The collected dataset for traffic flow and speed is filtered according to the Wright criterion, and the abnormal or missing data are estimated with the average of the same time interval on the same day of the past three weeks.

2) TIME DECOMPOSITION BASED ON SPEED-CLUSTERING

This study employs k-means cluster for travel speed time series to divide 24 hours of one day into multiple time periods in order to identify the peak and off-peak hours.

3) PERIOD-SPECIFIC PREDICTION FORMULATION

Based on the previous clustering results, the GM and BP model are combined for flow prediction at all divided time periods.

B. TIME PERIOD DECOMPOSITION BASED ON SPEED-CLUSTERING

As a partition-based clustering analysis, k-means algorithm has the advantage of efficiently processing huge dataset and discovering patterns. Based on the calculation of Euclidean distance, the objective function of clustering method can be expressed as follows:

$$SSE = \sum_{k=1}^{K} \sum_{v(i) \in C_k} dist(v(i), c_k)^2 \tag{2}$$

where, SSE means the summation of the squared error, which is regarded as objective function for clustering quality measuring; K is the total number of data clusters; C_k denotes the dataset of the kth cluster; $v(i)$ represents link travel speed at the ith hour in one day; and c_k is the centroid of cluster k.

Herein, this paper performed k-means clustering on travel speed to search for each clustering center. The entire procedure is decomposed into the following steps:

- Step 1: Initialize input variables $v(i)$ and set $K = 3$ based on our long-time field observation when we conducted over one-year signal timing optimization at the target signalization.

TABLE 1. Correlation analysis between current volume and previous interval ones on the same day.

Coefficients	$q(i\text{-}5)$	$q(i\text{-}4)$	$q(i\text{-}3)$	$q(i\text{-}2)$	$q(i\text{-}1)$	$q(i)$
$q(i\text{-}5)$	1	0.849	0.668	0.504	0.313	0.106
$q(i\text{-}4)$	0.849	1	0.850	0.669	0.507	0.315
$q(i\text{-}3)$	0.668	0.850	1	0.851	0.671	0.509
$q(i\text{-}2)$	0.504	0.669	0.851	1	0.851	0.672
$q(i\text{-}1)$	0.313	0.507	0.671	0.851	1	0.852
$q(i)$	0.106	0.315	0.509	0.672	0.852	1

TABLE 2. Correlation analysis between current volume and historical day ones at the same interval.

Coefficients	$q(i)$	$q(i\text{-}24)$	$q(i\text{-}48)$	$q(i\text{-}72)$	$q(i\text{-}96)$	$q(i\text{-}120)$	$q(i\text{-}144)$
$q(i)$	1	0.922	0.881	0.877	0.863	0.853	0.881
$q(i\text{-}24)$	0.922	1	0.922	0.887	0.884	0.866	0.858
$q(i\text{-}48)$	0.881	0.922	1	0.924	0.890	0.884	0.868
$q(i\text{-}72)$	0.877	0.887	0.924	1	0.926	0.889	0.884
$q(i\text{-}96)$	0.863	0.884	0.890	0.926	1	0.926	0.891
$q(i\text{-}120)$	0.853	0.866	0.884	0.889	0.926	1	0.923
$q(i\text{-}144)$	0.881	0.858	0.868	0.884	0.891	0.923	1

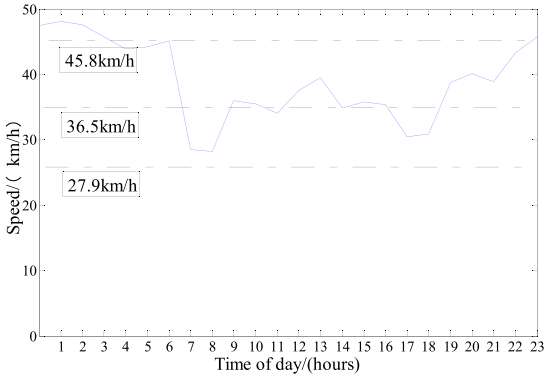


FIGURE 5. The clustering result for decomposed time periods.

- Step 2: Randomly choose K data samples of $v(j)$ as the initial cluster centers.
- Step 3: Calculate the distance from each sample in the dataset to all cluster centers, and then allocate this sample to the nearest center based on distance.
- Step 4: Update all cluster centers based on sample reallocation.
- Step 5: Repeat Step 3 and 4 until Equation (4) reaches the minimum, and obtain the final K clusters.

For the total 672 data samples, one can obtain the converged clusters after 10 iterations, and travel speed dataset are divided into three categories. If those samples for the same time interval on the different days belongs to more than one cluster, this study employs the majority voting method to tackle it. Finally, the clustering speed centers are 27.9km/h for peak period, 36.5km/h for flat-peak period, and 45.8km/h for low-peak period, respectively. Correspondently, the low-peak period ranges from 23:00 to 6:00, flat-peak period from 9:00 to 16:00 and 19:00 to 22:00, peak periods from 7:00 to 8:00 and 17:00 to 18:00 in Figure 5.

C. PERIOD-SPECIFIC PREDICTION MODEL

Based on the decomposed three periods in Figure 5, the period-specific combined predicted model (CPM) for 24-hour traffic flow is developed as follows:

1) GM-BASED PREDICTION FOR LOW-PEAK PERIOD

From after 23 to before 6 in one day, the average vehicle speed is close to the free-flow speed, and the traffic volume is very low. In particularly, the volume during this period is almost decreasing, so it's proper for GM model to capture this kind of downward trend with uncertainty. Actually, the grey theory model, the core component of the grey system, has been proved that it's been widely used in the field of transportation, especially for small-sample time series prediction or estimation [9]. Among, GM (1,1) is the typical format of grey theory, and can be formulated by the following Equations (3-7) [8]. Firstly, let's define the original time series as follows:

$$Q = (q(i+1), q(i+2), \cdots, q(i+d)) \tag{3}$$

And then, the one-time accumulated new series of traffic flow can be described by:

$$Q^{(1)} = (q^{(1)}(i+1), q^{(1)}(i+2), \cdots, q^{(1)}(i+d)) \tag{4}$$

where, $q^{(1)}(i+j) = \sum_{p=1}^{j} q(i+p)$, $1 \le j \le d$.

The superscript 1 means traffic flow is processed with accumulated generating operation from the original series. Subsequently, let's define the following expression:

$$\begin{cases} Z^{(1)} = (z^{(1)}(i+2), z^{(1)}(i+3), \cdots, z^{(1)}(i+d)) \\ z^{(1)}(m) = 0.5(q^{(1)}(m) + q^{(1)}(m-1)), 2 \le m \le i \end{cases} \tag{5}$$

where, $Z^{(1)}$ is a mean sequence of $Q^{(1)}$ calculated by formula $z^{(1)}(m) = 0.5(q^{(1)}(m) + q^{(1)}(m-1))$.And then, the basic GM(1,1) can be formulated by the following expressions:

$$q(m) + az^{(1)}(m) = b \tag{6}$$

where, a and b mean gray coefficients, which might be calibrated by the conventional statistical least-square method. Therefore, the predicted traffic flow can be expressed as follows:

$$q_{pre}(m+1) = q_{pre}^{(1)}(m+1) - q_{pre}^{(1)}(m) = (1-e^{a})(q(1) - \frac{b}{a})e^{-am}, \quad m \le i \quad (7)$$

Base on the previous correlation analysis in Section 2, this study took traffic flow time series in the past three hours as model inputs, and the output is the current hour one.

2) GM-BASED PREDICTION FOR FLAT-PEAK PERIOD

During the flat-peak period between 9:00 and 16:00, the average vehicle speed is medium compared with other two periods, but accompanied by a rapid rising or falling trend. Therefore, the GM is also suitable for it, and the inputs are the same as the low-peak period because the correlation coefficients of traffic flow between the current hour and the past three hours exceed 0.5.

3) BP-BASED PREDICTION FOR PEAK PERIOD

Artificial Neural Network can capture the fluctuation of traffic flow affected by the uncertain and nonlinear noises due to the capability of handling complex non-linear mapping, flexible network structure, and learning ability [46]. As a common neural network style, BP has the characteristics of signal forward transmission and error back propagation, and can capture the uncertainty and nonlinearity in traffic flow.

The remaining four hours belong to the peak hours ($j = 7$, 8, 17, and 18), and traffic volumes during peak hours have a greater fluctuation than others while they are much larger than others. Thus, BP neural network might be suitable to capture the fluctuation of traffic flow during peak hours. However, the data-driven BP model, a black box one, need much more sampling data to calibrate the parameters. In Table 2, this paper illustrated correlation analysis between the current hour volume and the same hour of the historical day of week is over 0.85, and thus employed historical data at the same hour belonging to the past serval weeks to train the BP model. The developed BP model has the popular structure of one input layer, one hidden layer, and one output layer, respectively.

As known, it's difficult to decide the number of neurons in the hidden layer for BP. According to the characteristics of the input and output data, the number of neurons in the hidden layer is initially determined by:

$$A = \sqrt{B + C} + D \quad (8)$$

where, B and C is the number of neurons in the input layer and output layer, respectively; and D denotes a constant integer from 0 to 10. After testing A values from 5 to 15, this study obtained the final value of 12 when the fitting error is the smallest. So, the structure of the network is BP(3,12,1), namely, 3 input neurons, 12 hidden ones and 1 output one, respectively. In this study, 504 of total 672 samples is selected as training ones, and the remaining is used for prediction.

Finally, the period-specific prediction model can be formulated in the following expression:

$$q'(i+1) = \begin{cases} f_{GM}(q(i-2), q(i-1), q(i)), 0 \le i \le 6 \\ f_{BP}(q(i-2), q(i-1), q(i)), i = 7 \text{ or } 8 \\ f_{GM}(q(i-2), q(i-1), q(i)), 9 \le i \le 16 \\ f_{BP}(q(i-2), q(i-1), q(i)), i = 17 \text{ or } 18 \\ f_{GM}(q(i-2), q(i-1), q(i)), 19 \le i \le 22 \\ f_{GM}(q(i-2), q(i-1), q(i)), i = 23 \end{cases} \quad (9)$$

where, $q'(i+1)$ is the predicted flow. If i-1 or i-2 is less than zero, it means the time series of the previous day will be regarded as inputs. For example, if $i = 0$, the $q(i\text{-}1)$ and $q(i\text{-}2)$ represent the volume at 23:00 and at 22:00 on the previous day, respectively.

IV. EXPERIMENTAL ILLUSTRATION

To evaluate the effectiveness of the proposed model, this study used the field data from Zhongshan for testing as shown in Section 2, and also compared with the popular existing models in terms of Measurement of Effectiveness (MOE) indexes as follows:

The Mean Absolute Error between the actual volume and predicted one

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|q(i) - q'(i)\right| \quad (10)$$

the Mean Absolute Percentage Error

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|q(i) - q'(i)\right|}{q(i)} \quad (11)$$

and the Mean Squared Error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(q(i) - q'(i)\right)^2 \quad (12)$$

where, n is the total number of tested data samples.

A. OVERALL PREDICTION ACCURACY

This study tested the proposed combined prediction model (CPM) and other nine models, including LSTM, ARIMA-like (ARIMA(0,0,12), BP-ARIMA(2,0,8), and GM-ARIMA(2,0,4)), BP-like (BP, ARIMA(0,0,12)-BP, and GM-BP), and GM-like (GM, ARIMA(0,0,12)-GM and BP-GM). Among, ARIMA-like means it chooses ARIMA as the underlying model of integrating other methods. Meanwhile, the six combined models, including ARIMA-BP (A-B), ARIMA-GM (A-G), BP-ARIMA (B-A), BP-GM (B-G), GM-ARIMA (G-A) and GM-BP (G-B) are implemented when the prediction from the former model (e.g. ARIMA) is regarded as one input of the latter one (e.g. BP), namely combination

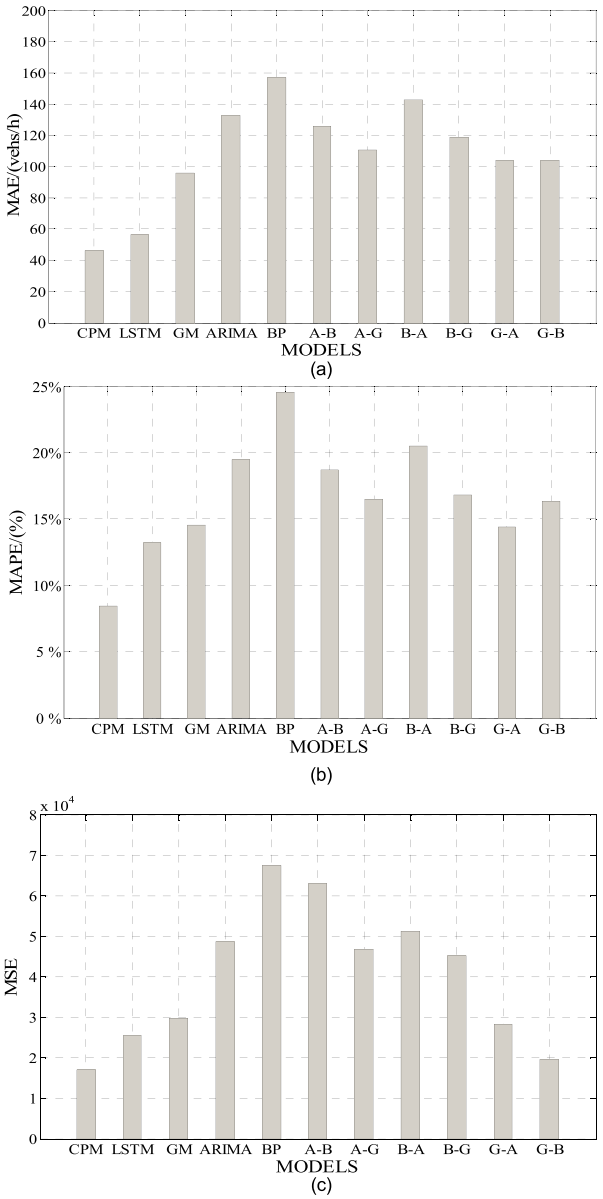


FIGURE 6. MOEs under 10 prediction models (a) MAE; (b) MAPE; (c) MSE.

(e.g. ARIMA-BP or A-B). Notably, in order to fairly evaluate these models, the related GM and BP has the same underlying structure of GM(1,1) and BP(3,12,1), and the parameters of all models will be retrained according to the sampled dataset.

The comparison of the prediction accuracy is shown in Figure 6. A conclusion can be reached that the proposed combination in this paper is capable to obtain much better results than others with regardless of MOEs. Among, the MAPEs of BP and BP-ARIMA exceed 20%, and the MAPE of other models except CPM ranges from 13.2% to 19.5%. Moreover, the performance of GM model and GM-ARIMA and GM-BP with the similar fundamental inputs from GM prediction is better than that of BP or ARIM models because the latter methods might not capture the upward or downward trend of traffic flow over time and more easily converge to a local

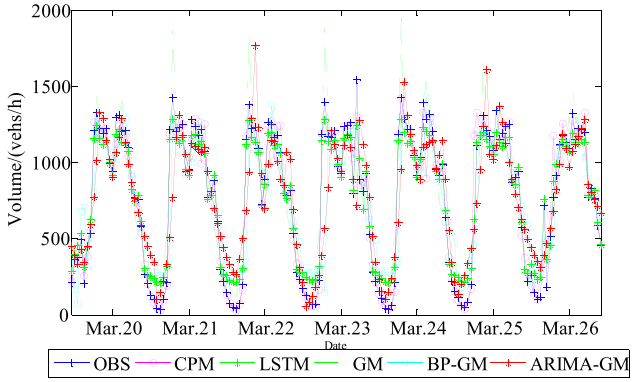


FIGURE 7. Comparison of CPM and GM-like model.

optimum. Notably, LSTM performs better than those existing ARIMA-like BP-like and GM-like models, but a litter worse than the period-specific CPM developed in this study. Overall, the GM model provides a good prediction accuracy, and thus this study presented GM and BP combined methods to separately predict traffic flow for different time of day. The results show that the MAPE of CPM reduces to the lowest value of 8.5% than other ten models.

B. CPM VS GM-LIKE MODELS OVER TIME

As shown in Figure 7, the GM model has a strong ability to track the fast descending or ascending trends of traffic volume, and show a much better accuracy than others. For example, from 10:00 to 12:00 on March 24, the actual flow by hour is about 1217 vehs/h, 1218 vehs/h, and 1043 vehs/h, respectively; and the corresponding prediction from GM model is 1219 vehs/h, 1128 vehs/h, and 1044 vehs/h, respectively. However, for the peak-hour period at 7:00 or 17:00, the low-peak one between 3:00 and 4:00 and the stationary period at 14:00 and 15:00, the overall prediction precision is not acceptable due to the overfitting problem of GM. On the contrary, the proposed period-specific combination model can suppress the overfitting for the dramatic flow fluctuation by importing BP.

The ARIMA-GM and BP-GM model have a little tracking ability and high accuracy during the periods when traffic flow increases or decreases rapidly because the underlying model is GM, which is similar to that of the single GM model. After integrating the ARIMA into GM, the overall forecasting accuracy shows a steady trend, and the overfitting problems for the maximal and minimal traffic flow prediction are improved compared to the single GM one. After integrating the BP model into GM, the overall prediction accuracy is generally improved, but the problem of time-lag occurs. Further, the LSTM has an under-fitting problem during the wave and valley peaks.

In details, the ARIMA-GM model shows a low prediction accuracy with the largest error of 40.7% during morning peak hours from Tuesday to Saturday, especially for 7:00 on March 24 (Friday). What's more, the proposed CPM with GM

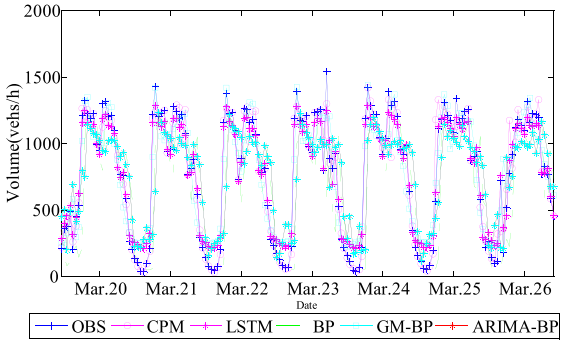


FIGURE 8. Comparison of CPM and BP model.

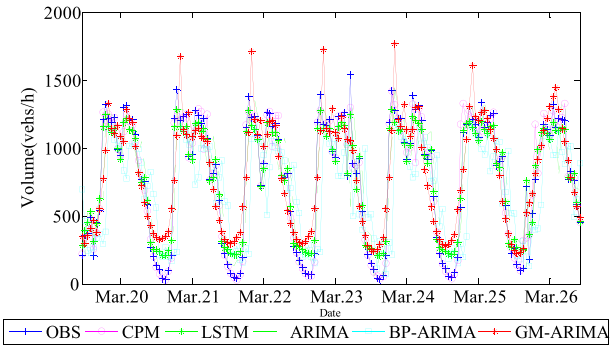


FIGURE 9. Comparison of CPM and ARIMA model.

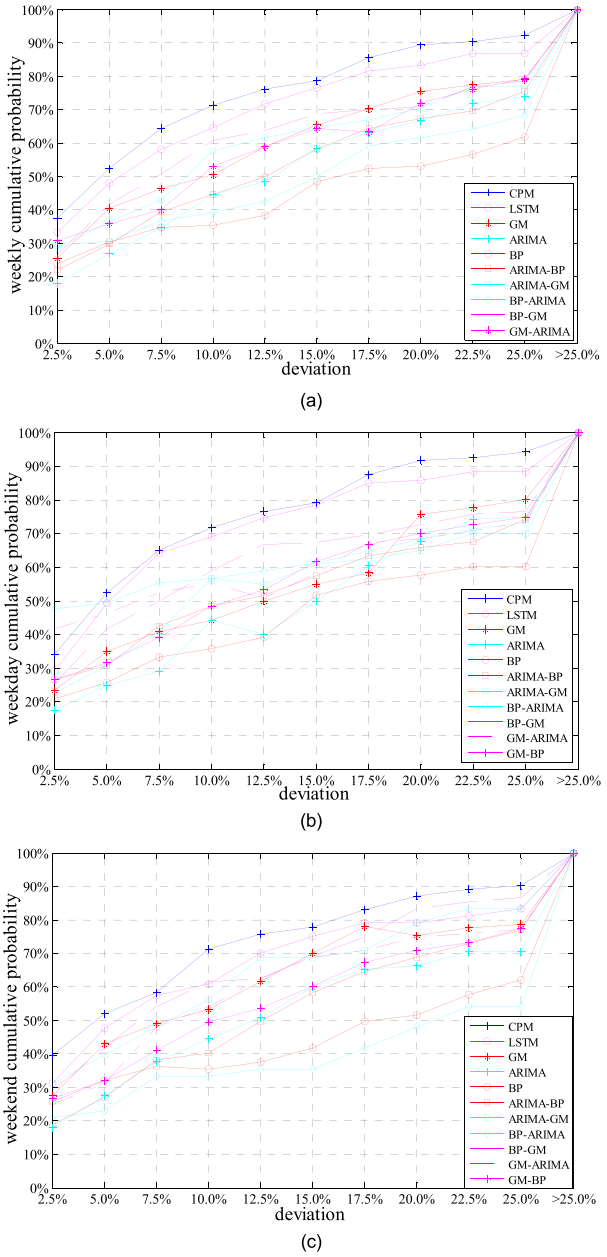


FIGURE 10. Cumulative probability density function of MAPE of 10 models: (a)one-week CDF; (b)weekday CDF; (c)weekend CDF.

can yield much better accuracy than the other two models (ARIMA-GM and BP-GM), especially the predication error drops to 0.3% at 7:00 on Friday.

C. CPM VS BP-LIKE MODELS OVER TIME

As shown in Figure 8, the performance of individual BP model is better than GM and ARIMA models during morning peak hours between 7:00 and 9:00 and evening peak hours between 17:00 and 19:00. However, during the periods of flow increasing or decreasing process, it shows lower accuracy and has time-lag characteristic of about one hour. For example, the hourly traffic flow from 4:00 to 8:00 on March 22 are 41 vehs/h, 75 vehs/h, 196 vehs/h, 1152 vehs/h, respectively, and the corresponding predicted one by BP is 50 vehs/h, 53 vehs/h, 891 vehs/h, and 1205 vehs/h, respectively. However, by importing other methods into BP, the ARIMA-BP and GM-BP models have the low performance with a prediction error larger than 30% during low-peak hours from 0:00 to 5:00. Compared with Figure 7, the same finding of LSTM in can be reached in Figure 8.

D. CPM VS ARIMA-LIKE MODELS OVER TIME

The ARIMA model provides a quite stable prediction accuracy over time as shown in Figure 9. However, the characteristic of time-lag by ARIMA is very obvious between 3:00 and 6:00 every day. Meanwhile, there is a weak tracking ability at serval specific periods (after 22:00 before 7:00, and 19:00-21:00) when the volume has the significant upward or downward trend, where the forecasting value is much smaller than observed one. The main reason is that the ARIMA is only capable to help understand the linear and stationary relationship of data and cannot capture the real-time fluctuation of traffic flow. The MAPE of LSTM (13.2%) is much better than that of GM-ARIMA (14.4%) and BP-ARIMA (20.5%). However, during peak-hour periods when traffic flow reaches to the maximum, the error of GM-ARIMA is greatly higher than that of BP-ARIMA because of the overfitting of the GM.

E. RELIABILITY ANALYSIS OF 10 MODELS

From the cumulative probability density function (CDF) curve of the prediction error under the proposed CPM and

other nine models as shown in Figure 10, the period-specific CPM model shows promising in prediction accuracy and reliability with regardless of time of day. Particularly, the probability with the MAPE of less than 10% in a week is up to 71.4% by CPM, while the probability of having larger than 25% MAPE is no greater than 5.8% on weekdays and 9.7% at weekends, respectively. Notably, the LSTM have a potential to achieve a much better prediction performance than other traditional methods except CPM.

V. DISCUSSIONS AND CONCLUSION

The objective of this study is to optimize a traditional 24-hour prediction logic and thereby develop a novel time-decomposition prediction method according to the fluctuation of traffic flow over time. In a test case, the field traffic volume and link travel speed with the interval of 1 hour were collected in Zhongshan. Firstly, according to the Wright criterion, the abnormal and missing data were processed. And then, temporal correlation analysis was performed to prove that the current traffic parameters have a significant correlation with those of the last three hours and the same time of day in the past seven days. After that, cluster analysis was conducted based on link travel speed, and the 24-hour time was divided into three periods, namely peak flow, flat peak flow and off-peak flow. Finally, period-specific prediction method with gray theory method and BP artificial neural network was formulated based on the characteristics of each divided period. A comprehensive experiment was conducted to validate the developed model, and it is found out that the mean absolute percentage error is about 8.46%. Most importantly, the probability of MAPE no more than 10% is close to 71.66% on weekdays, and 71.30% at weekends, respectively. What's more, the other nine models (namely, ARIMA-like, BP-like and GM-like and LSTM) are evaluated and compared in terms of MOEs and reliability, which can offer valuable insight at the field of both academia and industry. The proposed combined method can provide reliable information for traffic police departments on signal timing optimization and traffic guidance. The future work will focus on how to avoid the over-fitting and under-fitting of predicted models for the specific period.

ACKNOWLEDGMENT

The authors especially appreciate data support from Zhongshan Traffic Police Detachment, China.

REFERENCES

[1] B. Yao, C. Chen, Q. Cao, L. Jin, M. Zhang, H. Zhu, and B. Yu, "Short-term traffic speed prediction for an urban corridor," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 32, no. 2, pp. 154–169, Feb. 2017.

[2] 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, pp. 3–19, Jun. 2014.

[3] 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.

[4] S. Sun, C. Zhang, and G. Yu, "A Bayesian network approach to traffic flow forecasting," *IEEE Trans. Intell. Transp. Syst.*, vol. 7, no. 1, pp. 124–132, Mar. 2006.

[5] S. V. Kumar and L. Vanajakshi, "Short-term traffic flow prediction using seasonal ARIMA model with limited input data," *Eur. Transp. Res. Rev.*, vol. 7, no. 3, p. 21, Sep. 2015.

[6] T. Zhou, D. Jiang, Z. Lin, G. Han, X. Xu, and J. Qin, "Hybrid dual Kalman filtering model for short-term traffic flow forecasting," *IET Intell. Transp. Syst.*, vol. 13, no. 6, pp. 1023–1032, Jun. 2019.

[7] L. L. Ojeda, A. Y. Kibangou, and C. C. de Wit, "Adaptive Kalman filtering for multi-step ahead traffic flow prediction," in *Proc. Amer. Control Conf.*, Washington, DC, USA, Jun. 2013, pp. 4724–4729.

[8] A. Bezuglov and G. Comert, "Short-term freeway traffic parameter prediction: Application of grey system theory models," *Expert Syst. Appl.*, vol. 62, pp. 284–292, Nov. 2016.

[9] S. Mao, X. Xiao, M. Gao, X. Wang, and Q. He, "Nonlinear fractional order grey model of urban traffic flow short-term prediction," *J. Grey Syst.*, vol. 30, no. 4, 2018.

[10] I. Okutani and Y. J. Stephanedes, "Dynamic prediction of traffic volume through Kalman filtering theory," *Transp. Res. B, Methodol.*, vol. 18, no. 1, pp. 1–11, Feb. 1984.

[11] H. Sun, H. X. Liu, H. Xiao, R. R. He, and B. Ran, "Use of local linear regression model for short-term traffic forecasting," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1836, no. 1, pp. 143–150, Jan. 2003.

[12] M. Dougherty, "A review of neural networks applied to transport," *Transp. Res. C, Emerg. Technol.*, vol. 3, no. 4, pp. 247–260, Aug. 1995.

[13] M. N. Postorino and G. M. L. Sarnè, "Mobility forecast in an urban area through the use of neural networks," *Appl. Adv. Technol. Transp. Eng.*, vol. 13, no. 1, pp. 213–217, Jun. 1995.

[14] R. More, A. Mugal, S. Rajgure, R. B. Adhao, and V. K. Pachghare, "Road traffic prediction and congestion control using artificial neural networks," in *Proc. Int. Conf. Comput., Anal. Secur. Trends (CAST)*, Dec. 2016, pp. 52–57.

[15] S. Clark, "Traffic prediction using multivariate nonparametric regression," *J. Transp. Eng.*, vol. 129, no. 2, pp. 161–168, Mar. 2003.

[16] Y. F. Tang, W. H. K. Lam, and P. L. P. Ng, "Comparison of four modeling techniques for short-term AADT forecasting in Hong Kong," *J. Transp. Eng.*, vol. 129, no. 3, pp. 271–277, May 2003.

[17] G. Zhang, B. E. Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks: The state of the art," *Int. J. Forecasting*, vol. 14, no. 1, pp. 35–62, Jul. 1998.

[18] X. R. Liang and Y. X. Wei, "Freeway traffic flow modeling based on recurrent neural network and wavelet transform," in *Proc. Int. Conf. Transp. Eng.*, Chengdu, China, Jul. 2007, pp. 1088–1093.

[19] M.-C. Tan, S. C. Wong, J.-M. Xu, Z.-R. Guan, and P. Zhang, "An aggregation approach to short-term traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 1, pp. 60–69, Mar. 2009.

[20] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, Apr. 2014.

[21] D. Yu, Y. Liu, and X. Yu, "A data grouping CNN algorithm for short-term traffic flow forecasting," in *Proc. Asia–Pacific Web Conf.*, Suzhou, China, Sep. 2016, pp. 92–103.

[22] Y. Xu, D. W. Hu, and B. Su, "Short-term traffic flow prediction based on optimised support vector regression," *Int. J. Appl. Decis. Sci.*, vol. 10, no. 4, pp. 305–314, Oct. 2017.

[23] X. Dai, R. Fu, E. Zhao, Z. Zhang, Y. Lin, F.-Y. Wang, and L. Li, "DeepTrend 2.0: A light-weighted multi-scale traffic prediction model using detrending," *Transp. Res. C, Emerg. Technol.*, vol. 103, pp. 142–157, Jun. 2019.

[24] 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, pp. 1–14, May 2019.

[25] 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.

[26] N. G. Polson and V. O. Sokolov, "Deep learning for short-term traffic flow prediction," *Transp. Res. C, Emerg. Technol.*, vol. 79, p. 117, Jun. 2017.

[27] X. Xiao and H. Duan, "A new grey model for traffic flow mechanics," *Eng. Appl. Artif. Intell.*, vol. 88, Feb. 2020, Art. no. 103350.

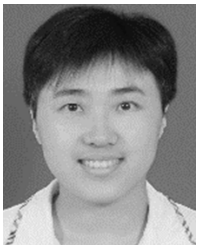
[28] K. C. P. Wang and Q. Li, "Pavement smoothness prediction based on fuzzy and gray theories," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 26, no. 1, pp. 69–76, Jan. 2017.

[29] M. Khashei and M. Bijari, "A novel hybridization of artificial neural networks and ARIMA models for time series forecasting," *Appl. Soft Comput.*, vol. 11, no. 2, pp. 2664–2675, Mar. 2011.
[30] Y. Chen, Y. Zhao, and P. Yan, "Daily ETC traffic flow time series prediction based on K-NN and BP neural network," in *Proc. Int. Conf. Pioneering Comput. Scientists, Eng. Educators*, Singapore, Aug. 2016, pp. 135–146.
[31] Y. Zhang and Y. C. Liu, "Application of combined forecasting models to intelligent transportation systems," in *Opportunities and Challenges for Next-Generation Applied Intelligence*, vol. 214. New York, NY, USA: Springer, 2009, pp. 181–186.
[32] J. Tang, L. Li, Z. Hu, and F. Liu, "Short-term traffic flow prediction considering spatio-temporal correlation: A hybrid model combing type-2 fuzzy C-means and artificial neural network," *IEEE Access*, vol. 7, pp. 101009–101018, Jul. 2019.
[33] 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.
[34] W. Feng, H. Chen, and Z. Zhang, "Short-term traffic flow prediction based on wavelet function and extreme learning machine," in *Proc. 8th IEEE Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Nov. 2017, pp. 531–535.
[35] Y. K. Wu, H. C. Tan, L. Q. Qin, B. Ran, and Z. X. Jiang, "A hybrid deep learning-based traffic flow prediction method and its understanding," *Transp. Res. C, Emerg. Technol.*, vol. 90, pp. 166–180, Apr. 2018.
[36] Z. Zheng, L. Pan, and K. Pholsena, "Mode decomposition based hybrid model for traffic flow prediction," in *Proc. IEEE 3rd Int. Conf. Data Sci. Cyberspace (DSC)*, Jun. 2018, pp. 521–526.
[37] B. Chang and H. Tsai, "Forecast approach using neural network adaptation to support vector regression grey model and generalized auto-regressive conditional heteroscedasticity," *Expert Syst. Appl.*, vol. 34, no. 2, pp. 925–934, Feb. 2008.
[38] Z. Ma and W. Liu, "Outlier correction method of telemetry data based on wavelet transformation and wright criterion," *Multimedia Tools Appl.*, vol. 75, no. 22, pp. 14477–14489, Nov. 2016.
[39] F. Moretti, S. Pizzuti, S. Panzieri, and M. Annunziato, "Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling," *Neurocomputing*, vol. 167, pp. 3–7, Nov. 2015.
[40] A. Ermagun and D. Levinson, "Spatiotemporal short-term traffic forecasting using the network weight matrix and systematic detrending," *Transp. Res. C, Emerg. Technol.*, vol. 104, pp. 38–52, Jul. 2019.
[41] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, and Y. Wang, "Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction," *Sensors*, vol. 17, no. 4, p. 818, 2017.
[42] M. T. Asif, J. Dauwels, C. Yang Goh, A. Oran, E. Fathi, M. Xu, M. M. Dhanya, N. Mitrovic, and P. Jaillet, "Spatiotemporal patterns in large-scale traffic speed prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 794–804, Apr. 2014.
[43] C. Antoniou, H. N. Koutsopoulos, and G. Yannis, "Dynamic data-driven local traffic state estimation and prediction," *Transp. Res. C, Emerg. Technol.*, vol. 34, pp. 89–107, Sep. 2013.
[44] S. Yang, "On feature selection for traffic congestion prediction," *Transp. Res. C, Emerg. Technol.*, vol. 26, pp. 160–169, Jan. 2013.
[45] M. Althoff, D. Hess, and F. Gambert, "Road occupancy prediction of traffic participants," in *Proc. 16th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Hague, The Netherlands, Oct. 2013, pp. 99–105.
[46] 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.

BIN FENG received the master's degree from the School of Electro-Mechanical Engineering, Guangdong University of Technology, China, in 2014. He is currently pursuing the Ph.D. degree with the School of Civil Engineering and Transportation, South China University of Technology, China. He has published four journal and conference academic articles. His current research interests include traffic signal control and traffic prediction.

JIANMIN XU received the B.S. degree in electrical engineering and automation from the Jiangxi University of Engineering, in 1982, and the M.S. degree in electrical engineering and automation and the Ph.D. degree in control theory and control engineering from the South China University of Technology, China, in 1986 and 1994, respectively. He is currently a Professor with the School of Civil Engineering and Transportation, South China University of Technology. He has published more than 70 academic articles. His main research interests include traffic information engineering and control, control theory, and control engineering. His research of traffic signal control system based on sub-area dynamic division and coordinated interaction technology has received the second prize of the 2014 Guangdong Science and Technology Progress Award.

YONGJIE LIN (Member, IEEE) received the Ph.D. degree from the School of Control Science and Engineering, Shandong University, China, in 2014. He was with the University of Maryland, College Park, USA, for a period of one year, as a Visiting Student, which was supported by the China Scholarship Council, China. He held a one-year postdoctoral position with the Department of Civil and Environmental Engineering, Northwestern University, USA. He is currently an Assistant Professor with the Department of Transportation, South China University of Technology, China. He has published over 30 journal and conference academic articles. His current research interests include developing traffic data analysis and mining, traffic signal control, and traffic prediction and simulation.



PENGHAO LI is currently a Deputy Detachment Leader with the Traffic Police Detachment, Zhongshan. She has been engaged in urban road traffic management and control for many years. She is currently responsible for collecting intelligent traffic collection data, business data and internet data, using statistical analysis, data mining, artificial intelligence, and other technologies to achieve the goal of traffic management from experience governance to data governance.