Modern Physics Letters B
Vol. 35, No. 13 (2021) 2150212 (16 pages)
© World Scientific Publishing Company
DOI: 10.1142/S0217984921502122

World Scientific
www.worldscientific.com

A hybrid approach for short-term traffic flow forecasting based on similarity identification

Wenjun Li* and Si Chen†

School of Economics and Management,
Jiangsu University of Science and Technology,
Zhengjiang 212003, China
**liwenjun@bjtu.edu.cn*
†1472703958@qq.com

Xiaoquan Wang‡

School of Traffic and Transportation,
Beijing Jiaotong University, Beijing 100044, China
15120886@bjtu.edu.cn

Chaoying Yin

College of Automobile and Traffic Engineering,
Nanjing Forestry University, Nanjing 210037, China
cyyin@126.com

Zhaoguo Huang

School of Civil Engineering,
Lanzhou University of Technology, Lanzhou 730050, China
seuhuang369@hotmail.com

Received 4 September 2020
Revised 11 November 2020
Accepted 13 November 2020
Published 18 February 2021

Short-term traffic flow forecasting is a key component of intelligent transportation system, yet difficult to be forecasted reliably, and accurately. A novel hybrid forecasting model is proposed by combining three predictors, namely, the autoregressive integrated moving average (ARIMA), back propagation neural network (BPNN) and support vector regression (SVR). First, it is assumed that all previous intervals can have influence on the predicted interval and then the entropy-based gray relation analysis method is applied to analyze the correlation and determine the length of time constrain window. Second, an improved Euclidean distance is employed to identify the similarity. Furthermore, the rank-exponent method is utilized to rank the results according to the similarity and

‡Corresponding author.

fuse the predicted values of the predictors. Finally, a numerical experiment is implemented, which indicates that the performance of forecasting results is superior to the conventional ones.

Keywords: Traffic flow; short-term traffic forecasting; similarity identification; time constrain window; entropy-based gray relation analysis; rank-exponent method.

1. Introduction

The urban road transportation system is a complex dynamic system which is often involved and used in our daily life. This system has obvious time variability, uncertainty and complexity. In order to study the system, early scholars at the beginning used some statistical methods to study the statistical properties and found the fundamental diagram (FD) of traffic flow theory. Then, there are a lot of experts and scholars based on bottom-up modeling method such as cellular automata theory,[1–16] car-following model theory.[17–20] Under the background of big data, some scholars applied machine learning methods to the study of traffic flow problems.[21,22] At present, due to the advancement of wireless communication technology and the development of computation capacity and storage capacity of computer equipment, traffic flow prediction based on historical data has become a new promising method.

Short-term traffic flow forecasting (STTFF) is a crucial study in traffic and transportation system study, which can provide theoretical and data supports for traffic management strategies.[23] STTFF is of great significance in traffic operation and planning. For example, traffic control strategies, such as ramp metering, pricing and information provision, require precise traffic forecasting information in order to mitigate congestion effectively. Therefore, it is always of interest for scholars to develop methods to improve the forecasting accuracy. However, it is a challenge to develop accurate forecasting algorithm incorporating the nonlinearity and uncertainty of traffic flow.[24,25]

The estimation approach consists of methods that estimate the traffic state, based on *a priori* knowledge of traffic and partial observation. *A priori* knowledge (i.e. assumptions) could be physical traffic flow models, which is usually obtained by abstracting actual traffic by employing physical principles and statistical/machine-learning (ML) methods. The aforementioned models have been extensively applied for model-driven traffic state estimation (TSE). The Lighthill, Whitham and Richards (LWR) model is the most widely applied.[26] Some studies have used part of the LWR model's implications (i.e. propagation speed of a shockwave) for TSE.[27] Besides, higher-order models are used in TSE studies to estimate the nonequilibrium traffic state and/or acceleration rates, which are significantly important for some traffic operations and ignored by the LWR model. For example, the Predictive-Wakeup (PW) model and its extension,[28,29] higher-order model modified from the LWR model, Aw-Rascle-Zhang (ARZ) model and its extension[30] and several Parallel Traffic Management Systems (PTMs) have been used.[31,32] Actual traffic is often multi-lane and multi-class; however, the models aforementioned are based on single-lane and single-class. Some linking models that explicitly consider multi-lane and/or multi-class behavior of traffic have been developed. These extensions are quite advanced and technical topics; therefore, refer, e.g. to van Wageningen-Kessels *et al.*[31]; and references therein for the case of the LWR model. One of the main challenges for such extensions is to model lane-changing behavior, in which human behavior is essentially important. Nevertheless, it should be noted that the ARZ model and its extension can be considered as a multi-class LWR model because it can be interpreted as the LWR model with vehicle-specific FDs.

Another important aspect of traffic flow modeling is the stochastic factors. Stochastic models have been developed to explain the uncertainty of traffic dynamics and effect of the input data's uncertainty on the model's output. For example, Sumalee *et al.*[32] developed a stochastic model based on the LWR model that considered both factors. In the model of Jabari *et al.*,[33] the source of stochasticity in traffic dynamics accounted for vehicle heterogeneity, which implied a relation between multi-class and stochastic models. In addition to the above model-driven forecasting methods, data-driven forecasting methods have also been greatly developed.

In past decades, numerous data-driven forecasting models have been developed to forecast the traffic flow over short-term horizons, including the autoregressive integrated moving average (ARIMA) model,[34,35] Markov chain model,[36,37] Kalman filtering model,[38–40] the wavelet analysis-based model, support vector machine (SVM),[41] neural networks,[42–44] support vector regression (SVR)[45] and K-nearest neighbor algorithm.[46–48] The methods mentioned above are all proved to be powerful tools when they are utilized to forecast traffic flow. Each model demonstrates that satisfactory performance can be obtained for a certain traffic pattern because they have their own advantages.[49] Nevertheless, the ability to timely, reliably and accurately forecast the dynamics of traffic is difficult to meet by a single model in all applications. Owing to the drawbacks of each model, hybrid models combining multiple methods gradually emerge, which have been proved to be superior to individual ones by a few scholars.[42,49–57] For instance, Wang proposed a novel Bayesian combination method to fuse three predictors including ARIMA, BPNN and Kalman models in which BCM was improved by assuming that the transfer probability can be influenced by the errors of three predictors at all past intervals.[49] Tang combined the double exponential smoothing method and SVM to present the linear structure and nonlinear structure respectively, which outperformed the comparative methods selected in the study.[50] Li employed four different combinations of machine learning and time series forecasting models.[54] Wang implemented a hybrid ARIMA-EGARCH-M-GED model in which ARIMA model and EGARCH-M model were utilized to fit the linear part and error series, respectively.[57]

In this study, a novel hybrid approach is proposed that combines ARIMA, BPNN and SVR to take advantage of their own applicability under different traffic states because it is difficult to obtain precise predicting results using linear or nonlinear models separately.[54] In the combination model, ARIMA is used to capture the linear characteristics of traffic flow, and BPNN and SVR are used to capture the nonlinear parts. First, it is assumed that the traffic flow at a particular interval is influenced by ones at all previous intervals and the entropy-based gray relation analysis method is utilized to determine the length of time constrain window according to the correlation. Then an enhanced K-nearest neighbor algorithm is utilized to identify similarity measured by an improved Euclidean distance between observed time series and predictors, which is utilized as a reference for a progressive forecasting further. What is more, we rank the predictors according to the similarity and the rank-exponent method is used to fuse the results of predictors. Finally, the results indicate that the proposed model is superior to all individual methods.

As opposed to the separate parametric prediction or data-driven models, the method developed in this study can take advantage of the candidate values from different models. Moreover, this method is capable of accommodating different models by replacing the basic predicting models. In other words, the basic predicting models used in this study can be replaced with other predicting models, which can provide reliable predicting results.

This paper is organized as follows. Section 2 presents the methodology, including the framework, description of predictors, identification of similarity, time constrain window, rank-exponent based fusion model and evaluation method of forecasting results. Section 3 implements the experiments to evaluate the performance of the proposed method. Section 4 concludes this study and discusses the future work.

2. Methodology

The proposed algorithm aims to improve the forecasting results by exploiting the similarity between observed traffic flow and predicted ones in time constrain window in which the rank-exponent method is utilized to fuse the results from each predictor. The algorithm is shown in Fig. 1.

As shown in Fig. 1, the framework is summarized as follows:

Step 1: Predictor training and forecasting. The short-term historical traffic flow data is used to calibrate the parameters of the predictors and forecast the traffic flow according to the training results.

Step 2: Similarity identification. Entropy-based gray relation analysis method is utilized to determine the length of time constrain window and then exploit the similarity between observed traffic flow and predicted ones in time constrain window. According to the similarity of each predictor, we rank the value of each predictor.

Step 3: Fusion and forecasting. To obtain the final forecasting result, the rank-exponent method is used to fuse the forecasting result of each predictor, which is based on the ranking results.

2.1. *Description of predictors*

2.1.1. *ARIMA model*

ARIMA model is one of the most popular models for forecasting.[58] ARIMA(p, d, q) model is a mathematical model based on ARMA model, which is developed to

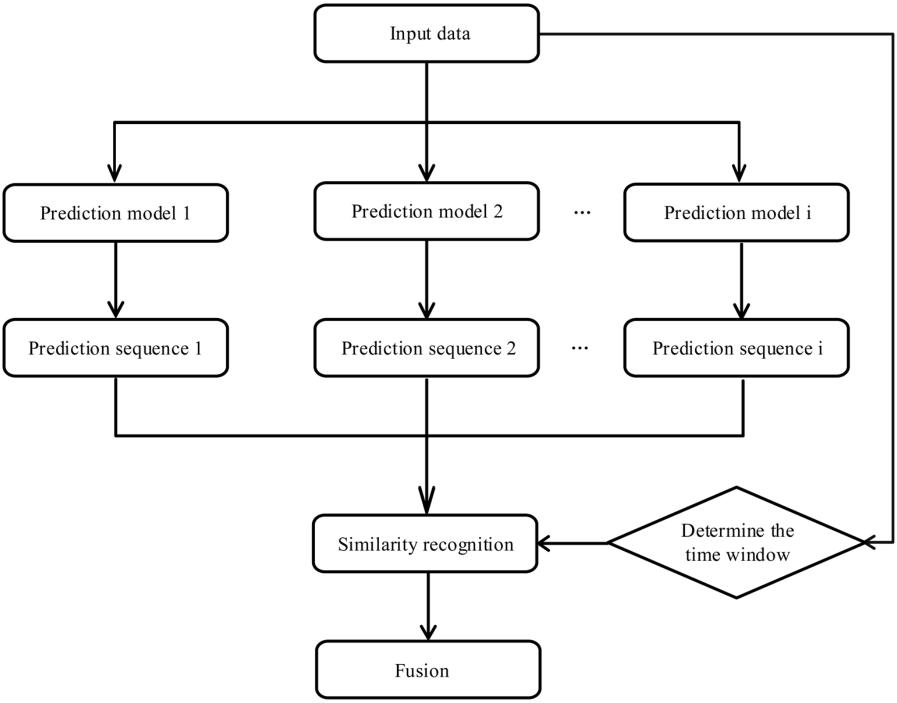


Fig. 1. Flow chart of the proposed method.

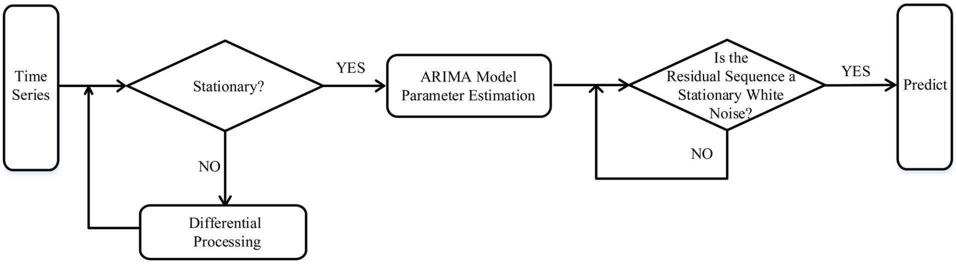

Time Series
Stationary?
YES
NO
Differential Processing
ARIMA Model Parameter Estimation
Is the Residual Sequence a Stationary White Noise?
YES
NO
Predict

Fig. 2. Implementation of the ARIMA model.

fit time series. In the ARIMA model, p, d, q represent the orders of AR, difference and MA, respectively. The model can be expressed as follows:

$$\varphi(B)(1-B)^d \widehat{y_t} = \theta(B)\varepsilon_t, \tag{1}$$

where $\widehat{y_t}$ is the traffic flow to be forecasted at tth interval. B is the backshift operator. ε_t is the white noise sequence, following $N(0, \sigma^2)$.

ARIMA model can be specifically implemented as follows:

Step 1: The stationarity of the time series is identified according to the AR and MA. If the time series is not stable, then we difference the time series until it become stable.

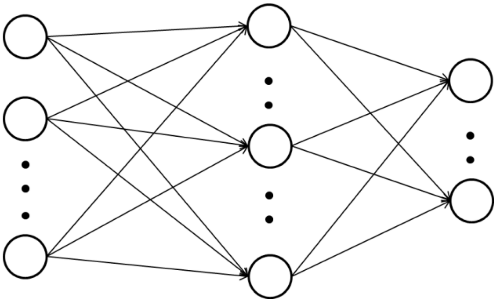


Fig. 3. Structure of the BPNN network.

Step 2: The maximum likelihood method is used to estimate the parameters in ARIMA model.

Step 3: We should judge whether the residual sequence is a white noise sequence. If it is a white noise sequence, go to Step 4. Otherwise, we estimate the parameters until the residual sequence becomes a white noise sequence.

Step 4: The estimated parameters are used to forecast the time series.

2.1.2. *Back propagation neural network*

The back propagation neural network (BPNN) model is a typical nonparametric model, which has been used in traffic flow forecasting universally. The structure of BPNN network includes the input layer, hidden layer and output layer, as shown in Fig. 3.

Before using the BPNN network to forecast the traffic flow, the historical traffic flow data is used to train the network so that it can arrange the weights of hidden layer. One of the characters of BPNN network is that information is spread along the positive direction and yet the error spreads oppositely during the training process. BPNN trains continuously until the error can meet precision requirements.

2.1.3. *Support vector regression model*

The forecasting model based on SVR is one of the most used intelligent models, which can model complex nonlinear dynamic systems. SVR model can convert the nonlinear programming problems into linear programming problems with the kernel function so that it can obtain the global optimal solution. Owing to the good generalization ability, we choose SVR model as a component predictor to model the nonlinear given observations and values to be forecasted. SVR model can solve the nonlinear programming problems through mapping inputs in low dimension into a higher Hilbert space. SVR model can be presented as follows:

$$f(x) = \omega^T \varphi(x) + b \,, \tag{2}$$

where $\omega = \{\omega_1, \omega_1, \ldots, \omega_D\}$ is the weights arranged b is the offset.

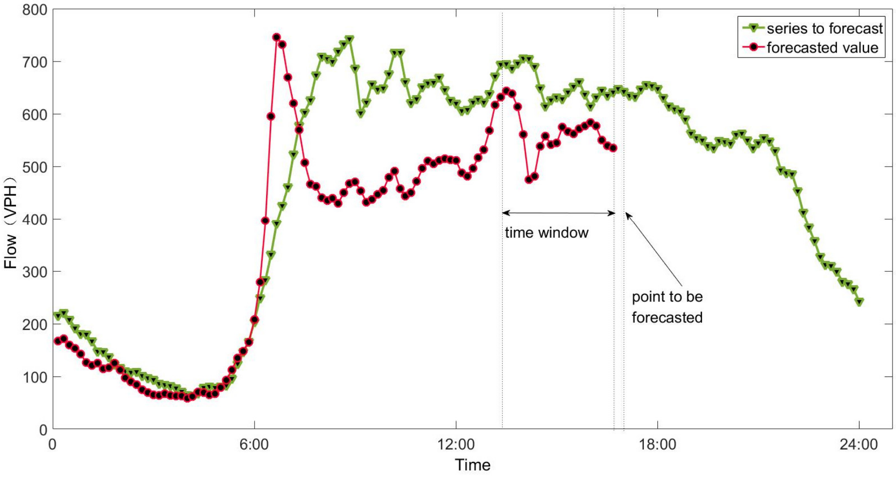


Fig. 4. (Color online) Similarity between the target sequence and a candidate sequence.

2.2. *Identification of similarity*

In the study, we forecast the traffic flow based on the similarity. The traffic flow at several intervals before the predicted interval is used to measure the similarity between the targeted sequence and candidate sequences from the predictors. Then the traffic flow value at the forecast interval is calibrated by fusing the values from candidate sequences according the similarity. The similarity between the target sequence and a candidate sequence can be calibrated, as shown in Fig. 4.

The definition of similarity aims to identify the distance between the targeted sequence and a candidate sequence, which presents the degree of proximity. Since the traffic flow tends to be related to the ones several intervals before to some extent, it is reasonable to arrange weights for the values of different predictors according to recentness of the values in the time constrain window.

2.2.1. *Linear weighted Euclidean distance*

There are several measurements of similarity of time series, including correlation distance, Euclidean distance, Gaussian weighted Euclidean distance and linear weighted Euclidean distance, which have been proved to be effective in pattern recognition for time series data.

Euclidean distance is a common measurement for the distance of two points. For time series x^m and y^m, we can obtain the distance between them by aggregating the distance between the related points in the time series. The distance can be presented as follows:

$$D_{(x^m,y^m)} = \sqrt{\sum_{i=0}^{m-1} (x_{T-i} - y_{T-i})^2}\,. \tag{3}$$

The traffic flow which is closer to the forecasting interval influences the predicted once more. In order to reflect the essential of the closeness, we arrange weights for each data point in the time series according to the closeness. The distance can be calibrated as follows:

$$D_{(x^m,y^m)} = \sqrt{\sum_{i=0}^{m-1} \omega_i (x_{T-i} - y_{T-i})^2}\,, \tag{4}$$

where ω_i is the weight associated with each point of the data. In this study, the weights are negatively associated with the recentness of the values and they are distributed linearly.

2.3. *Time constrain window*

It is important to determine the length of time constrain window because it should ensure that the data in the time constrain window is related to the forecasting one to some extent. It is a challenge to choose a suitable time constrain window so that rolling forecast can be used to raise forecasting accuracy. Because of the nonlinear characters of traffic flow, it is meaningless to choose a linear method to calibrate the correlation and use it to determine the length of time constrain window. In this study, the entropy-based gray relation analysis method is used to analyze the correlation so that we can determine the length of time window.

Denote that the traffic flow y_t at tth interval is influenced by its previous w-th intervals' traffic flows. $y_{t-1}, \dots, y_{t-w}, y_t$ and $y_{t-1}, \dots, y_{t-w}$ are presented by Y_t and $Y_{t-i} = \{y_{t-1}(j) | j \in J\}$, respectively, where Y_t is the target series and Y_{t-i} is the alternative traffic flow time series. Further, we denote $J = \{1, 2, \dots, M\}$ as the set of all forecasting values. Then the gray relational coefficient between them can be calibrated by the following equation:

$$\gamma(y_t(j), y_{t-i}(j)) = \frac{\min_i \min_j |y_t(j) - y_{t-i}(j)| + \zeta \max_i \max_j |y_t(j) - y_{t-i}(j)|}{|y_t(j) - y_{t-i}(j)| + \zeta \max_i \max_j |y_t(j) - y_{t-i}(j)|}\,, \tag{5}$$

where ζ is distinguishing coefficient, whose range is [0,1].

We can obtain the correlation between the target time series and the alternative traffic flow time series according to Eq. (5). Then, we map the target time series and the alternative traffic flow time series in order to obtain the mapping distribution presented as follows:

$$p(i,j) = \frac{\gamma(y_t(j), y_{t-i}(j))}{\sum_{j=1}^{M} \gamma(y_t(j), y_{t-i}(j))}\,. \tag{6}$$

According to the mapping distribution, the entropy of the gray relational coefficient can be calibrated as follows:

$$E(t-i) = \frac{-\sum_{j=1}^{M} p(i,j) \ln(p(i,j))}{\ln M}; \quad i = (1, 2, \dots, w)\,, \tag{7}$$

where $-\sum_{j=1}^{M} p(i,j)\ln(p(i,j))$ is the entropy of the gray relational coefficient, whose form is the same as Shannon information entropy. When $p(i,j)$ is equal, the gray correlation coefficient distribution is the most balanced and the gray entropy reaches the maximum value. The maximum value is only related to the number of attributes. We denote $\ln M$ as the maximum value of entropy of the gray relational coefficient. Note that the higher the entropy of the gray relational coefficient, the target time series is more related to the alternative traffic flow time series. However, the entropy of the gray relational coefficient and gray relational coefficient cannot reflect the correlation and balance of error distribution separately. We combine the entropy of the gray relational coefficient and gray relational coefficient so that the index can reflect both the characters. The entropy-based gray relation index is computational according to the following equation:

$$B(t-i) = \frac{E(t-i)}{M} \sum_{j=1}^{M} \gamma(y_t(j), y_{t-i}(j)) \,, \tag{8}$$

where $B(t-i)$ is the entropy-based gray relation index. It can be seen that the higher the entropy-based gray relation index, the target time series is more related to the alternative traffic flow time series. Then we can choose a threshold to ensure constrain time window in which the index higher than the threshold is determined to be contained.

2.4. *Rank-exponent method-based forecasting model*

It is important to assign weights for the predictor values to obtain accuracy predicting results because the influence of the predictors on the forecasted value are adjusted according to their proximity to the subject profile. The main assignment methods include: equal weights, inverse distance weights and rank-exponent method. Among three methods, the final one is found to have better performance in identifying similar traffic patterns.[57] The rank-exponent method is advantageous because we can adjust the assignment of weights flexibly by adjusting the weight dispersion measure. Therefore, the rank-exponent method is utilized to fuse the values of three predictors to improve the forecasting accuracy.

Specifically, the values of three predictors are ranked according to the similarity in the constrain time window. Then weights are arranged for predictors. The weights can be determined by the following equation:

$$W^i = \frac{(K - r_i + 1)^z}{\sum_{j=1}^{K}(K - r_i + 1)^z} \,, \tag{9}$$

where r_i is the rank of the ith predictor. K is the number of the predictors, and Z is the adjustment factor.

At the t-th interval, the forecasting value is the linear combination of the predictors, which is presented as follows:

$$\hat{y}_T = \sum_{n=1}^{N} W_{T-1}^{n} y_T^{n} \Bigg/ \sum_{n=1}^{N} W_{T-1}^{n}\,, \tag{10}$$

where $\hat{y}_T$ is the final forecasting value at the i-th interval, and y_T^n is a forecasting value of n-th predictor.

2.5. *Performance of forecasting result*

To test performance of the proposed approach, two indices are used to evaluate the performance of forecasting result: mean absolute percentage error (MAPE) and root mean square error (RMSE). The performance indices are defined as follows:

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \widehat{y_i}}{y_i}\right|, \tag{11}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \widehat{y_i})^2}{N}}\,, \tag{12}$$

where y_i and $\widehat{y_i}$ are the observed traffic flow and forecasting values, respectively, and N is the number of the forecasting values.

3. Numerical Study

In order to evaluate the performance of the proposed method, a numerical experiment is implemented based on the traffic flow data collected from Beijing expressway. The data is collected every 10 min. We divide the traffic flow data of four weekdays into two parts: data of three days is used to train the model and the data of the other day is used to forecast. The data to verify the model is collected from Beijing expressway Second Ring and Third Ring, respectively. The detectors are located as shown in Fig. 5.

Before implementing the proposed method, we estimate three predictors. The details are as follows:

ARIMA: Traffic flow data of three days is used to fit the model, and determine the orders of AR, differencing and MA. The final model is determined as ARIMA(7,1,1).

BPNN: Traffic flow data of three days is used to train the BP neural networks. The structure of BPNN contains: five neurons in the input layer, one hidden layer with ten neurons and one output neuron. The activation or transfer function in the hidden layer is chosen as the hyperbolic sigmoid function tanh.

SVR: The data of three days is utilized to train the $\varepsilon - SVR$ model. The genetic algorithm optimizes parameters of SVR model and the other day is used to forecast.

The forecasting results of three predictors are presented in Fig. 6 for data from detection located at Second Ring. It can be seen that the forecasting result of SVR

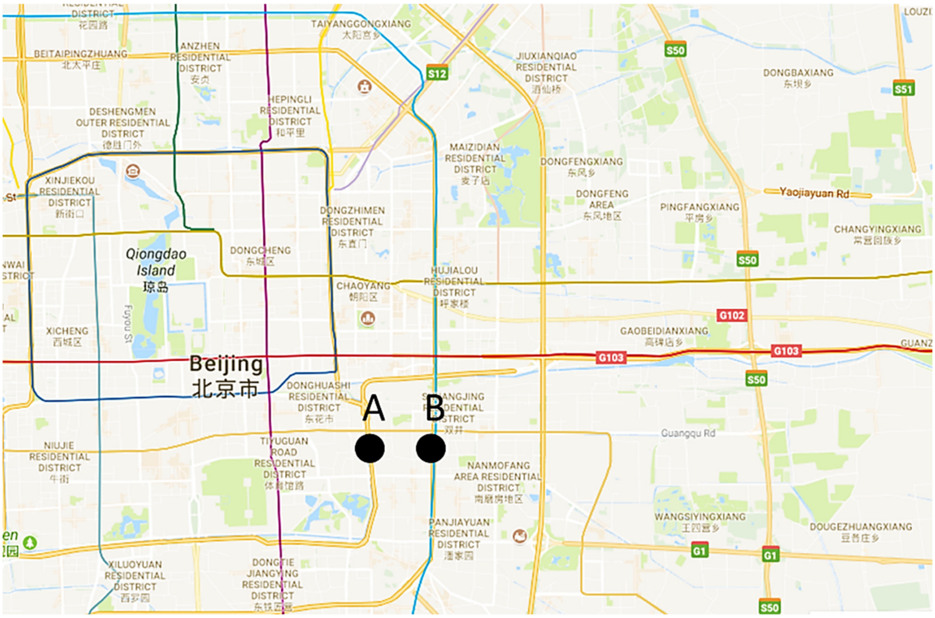

A
B
Beijing
北京市
DISTRICT
花园路
TAIYANGGONGXIANG
太阳宫乡
LOUZI
BEITAIPINGZHUANG
北太平庄
ANZHEN RESIDENTIAL DISTRICT
安贞
S12
S50
JIUXIANQIAO RESIDENTIAL DISTRICT
酒仙桥
DONGBAXIANG
东坝乡
S51
HEPINGLI RESIDENTIAL DISTRICT
和平里
DESHENGMEN OUTER RESIDENTIAL DISTRICT
德胜门外
MAIZIDIAN RESIDENTIAL DISTRICT
麦子店
DONGFENGXIANG
东风乡
XINJIEKOU RESIDENTIAL DISTRICT
新街口
DONGFENG AREA
东风地区
Yaojiayuan Rd
DONGZHIMEN RESIDENTIAL DISTRICT
东直门
PINGFANGXIANG
平房乡
CHANGYINGXIANG
常营回族乡
Qiongdao Island
琼岛
DONGCHENG
东城区
S50
Fuyou St
CHAOYANG
朝阳区
HUJIALOU RESIDENTIAL DISTRICT
呼家楼
G102
XICHENG
西城区
GAOBEIDIANXIANG
高碑店乡
G103
G103
GUANZ
S50
DONGHUASHI RESIDENTIAL DISTRICT
东花市
SHUANGJING RESIDENTIAL DISTRICT
双井
Guangqu Rd
NIUJIE RESIDENTIAL DISTRICT
牛街
TIYUGUAN ROAD RESIDENTIAL DISTRICT
体育馆路
NANMOFANG AREA RESIDENTIAL DISTRICT
南磨房地区
WANGSIYINGXIANG
王四营乡
DOUGEZHUANGXIANG
豆各庄乡
PANJIAYUAN RESIDENTIAL DISTRICT
潘家园
G1
G1
XILUOYUAN RESIDENTIAL DISTRICT
西罗园
DONGTIEJIANGYING RESIDENTIAL DISTRICT
东铁营
S50

Fig. 5. (Color online) Data collection location at Second Ring and Third Ring.

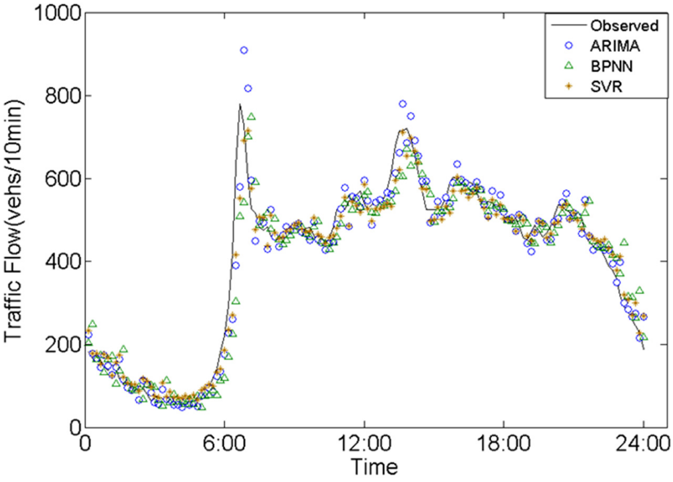


Fig. 6. (Color online) Forecasting results of three predictors.

performs better than ARIMA and BPNN model slightly. However, it is difficult for the performance of single predictor to meet the accuracy requirements.

It is essential to choose a suitable time constrain window for improving the forecasting accuracy. The entropy-based gray relation analysis method is utilized

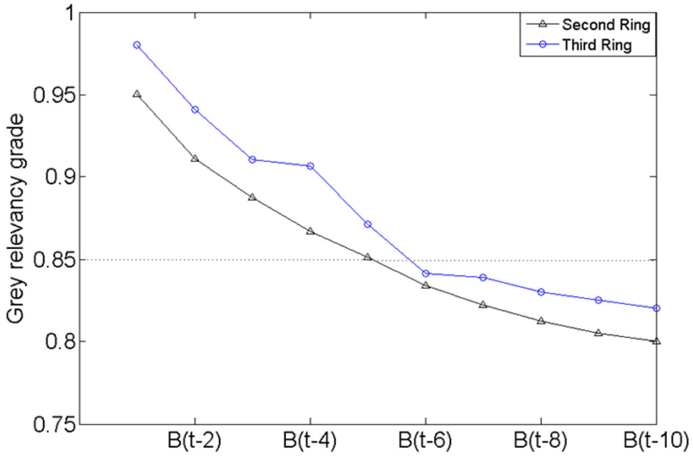


Fig. 7. (Color online) Entropy-based gray relation index at both locations.

to analyze the correlation between the target time series and the alternative traffic flow time series and determine the length of time constrain window further. Assume that the traffic flow at the predicted time has a strong correlation with the flow of the previous 10 moments, and the intervals of decisive influence on the forecasting time are selected. According to the entropy-based gray relation analysis method, the entropy-based gray relation index can be obtained at location A and B, which are presented in Fig. 7. Note that the correlation between two series decreases as the distance from the predicted time increases at both locations. The higher index demonstrates the interval is more related to the predicted interval. Therefore, we choose the intervals which are higher than the threshold to calibrate the similarity in time constrain window. In this study, we set the threshold to 0.85. Note that the length of time constrain window can be determined as 5 at both locations A and B, which means 50 min.

The similarity between observed time series and each predictor is utilized to rank the predictors and arrange weights for them, which are presented by Eq. (10). According to the rank-exponent method, the values of predictors are fused. Figure 8 shows the relationship between the variability of the result adjustment parameter Z and the forecasting accuracy of the proposed model with real data.

As shown in Fig. 8, when Z increases, MAPEs decrease substantially and then increase gradually with the fluctuations. Model accuracy is optimized when Z is equal to 0.03 and 0.12 at locations A and B, respectively.

In this study, the performance of proposed method is comprised with the single predictors at two locations. Table 1 presents the MAPE and RMSE values in the models. The proposed method relies on the predictors to some extent, which can improve the forecasted performance at both locations.

As presented in Table 1, the proposed method has a performance of 8.89% MAPE and 38.45 RMSE at the Second Ring, a 14.35% improvement for MAPE

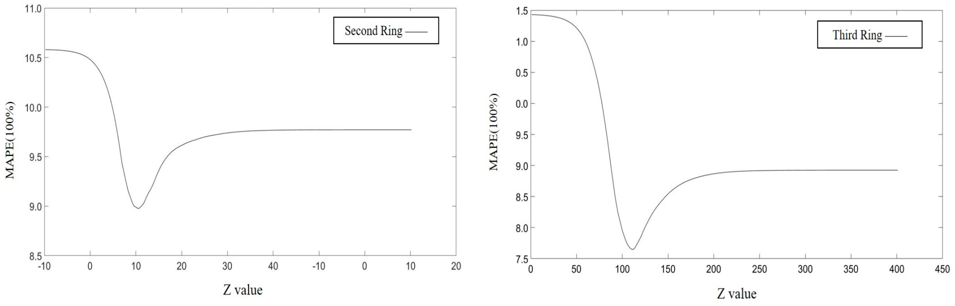
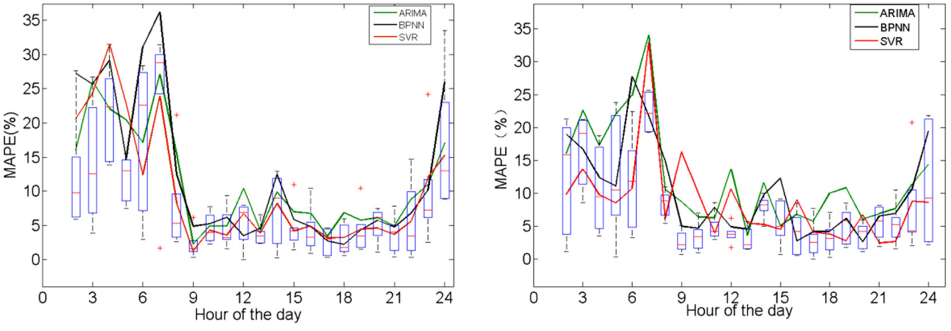


Fig. 8. Predictive performance under different values of the adjustment factor.

Table 1. Comprisation of different models.

	ARIMA		BP		SVR		KNN	
	MAPE (%)	RMSE	MAPE (%)	RMSE	MAPE (%)	RMSE	MAPE (%)	RMSE
Location A	11.48	56.12	12.64	61.81	10.38	44.03	8.89	38.45
Location B	12.16	73.09	11.07	50.01	11.44	57.54	7.65	32.71



(a) (b)

Fig. 9. (Color inline) Errors for proposed method and its comparison with component predictors.

and 12.67% for RMSE over the optimal one of three predictors. Correspondingly, the data collected at location B proves that the proposed method has a 30.89% improvement for MAPE and a 34.59% improvement for RMSE, respectively.

Figure 9 presents the forecasting errors by time of the day at Second Ring and Third Ring, respectively.

The box plots in Fig. 9 present the spread of the forecasting relative errors for the proposed method while the mean errors for three predictors are shown in solid green, black and red lines, respectively. Compared with the three predictors, the distribution of the errors for the proposed method is similar to them, which means that the forecasting accuracy of proposed method relies on the predictors to some extent. However, there is a statistically significant increase in mean forecasting errors. Due to the complexity of traffic state on peak time, the forecasting accuracy is unsatisfactory, yet improved to a greater extent.

4. Conclusions

This paper proposes a novel approach for forecasting traffic flow based on similarity identification. The entropy-based gray relation analysis is utilized to determine the length of time constrain window. Then rank-exponent method is used to fuse the results of ARIMA, BPNN and SVR according to the similarity patterns, measured by the linear weighted Euclidean distance in the window. The forecasting results show that the performance of proposed approach is satisfactory by using MAPE and RMSE as criteria. Due to the fact that all models have their own advantages and disadvantages in predicting short-term traffic flow (e.g. peak hours and off-peak hours), using a separate model to predict short-term traffic flow may not provide reliable predicting results. Therefore, in the scenarios which need real-time forecasting and updating, the method developed in this study can provide more reliable predicting results by combining different basic models.

One of the limitations of this study is that the forecasting accuracy of proposed method relies on the component predictors to some extent. In addition, the spatial characters are not considered in the proposed method. Future research should focus on further applying the proposed model for network level forecasting as well as improving the forecasting accuracy by improving the predictors.

References

1. A. K. Gupta and S. Sharma, *Chin. Phys. B* **21** (2012) 015201.
2. A. K. Gupta and I. Dhiman, *Int. J. Mod. Phys. C* **25** (2014) 1450045.
3. A. K. Gupta, S. Sharma and P. Redhu, *Commun. Theor. Phys.* **62** (2014) 393.
4. T. Nagatani, *Phys. A: Stat. Mech. Appl.* **419** (2015) 1.
5. T. Nagatani, *Phys. A: Stat. Mech. Appl.* **427** (2015) 92.
6. Y. Hino and T. Nagatani, *Phys. A: Stat. Mech. Appl.* **428** (2015) 416.
7. H. X. Ge, L. L. Lai, P. J. Zheng and R. J. Cheng, *Phys. Lett. A* **377** (2013) 3193.
8. H. X. Ge, P. J. Zheng, S. M. Lo and R. J. Cheng, *Nonlinear Dyn.* **76** (2014) 441.
9. W. X. Zhu and C. H. Zhang, *Phys. A: Stat. Mech. Appl.* **392** (2013) 3301.
10. L. D. Zhang, W. X. Zhu and J. L. Liu, *Phys. A: Stat. Mech. Appl.* **406** (2014) 89.
11. W. X. Zhu and R. L. Yu, *Phys. A: Stat. Mech. Appl.* **393** (2014) 101.
12. T. Q. Tang, K. W. Xu, S. C. Yang and C. Ding, *Phys. A: Stat. Mech. Appl.* **441** (2016) 221.
13. T. Q. Tang, W. F. Shi, H. Y. Shang and Y. P. Wang, *Nonlinear Dyn.* **76** (2014) 2017.
14. T. Q. Tang, W. F. Shi, S. C. Yang and H. Y. Shang, *Mod. Phys. Lett. B* **29** (2015) 1550020.
15. Y. S. Qian, X. L. Zhang, J. W. Zeng, X. M. Shao and N. Wang, *Mod. Phys. Lett. B* **29** (2015) 1450264.
16. Y. S. Qian, X. M. Shao, J. W. Zeng and M. Wang, *Phys. A: Stat. Mech. Appl.* **392** (2013) 5874.

17. H. Ou, T. Q. Tang, J. Zhang and J. M. Zhou, *Phys. A: Stat. Mech. Appl.* **505** (2018) 105.
18. P. Liao, T. Q. Tang, T. Wang and J. Zhang, *Phys. A: Stat. Mech. Appl.* **525** (2019) 108.
19. H. Ou and T. Q. Tang, *Phys. A: Stat. Mech. Appl.* **495** (2018) 260.
20. J. Zhang, T. Q. Tang and T. Wang, *IET Intel. Trans. Syst.* **13** (2019) 1686.
21. T. Q. Tang, Y. Gui, J. Zhang and T. Wang, *J. Transport. Eng. A: Syst.* **146** (2020) 04020104.
22. X. Wang, R. Jiang, L. Li, Y. L. Lin and F. Y. Wang, *Phys. A: Stat. Mech. Appl.* **514** (2019) 786.
23. D. W. Xia, B. F. Wang, H. Li and Z. L. Zhang, *Neurocomputing* **179** (2016) 246.
24. C. Dong, Z. Xiong, C. Shao and H. Zhang, *Transportmetric. A Transport Sci.* **11** (2015) 1.
25. H. Haj-Salem and J. Lebacque, *Transport. Res. Rec. J. Transport. Res. Board* **1802** (2002) 155.
26. M. Treiber and D. Helbing, *Cooperative Transport. Dyn.* **1** (2002) 1.
27. C. Nanthawichit, T. Nakatsuji and H. Suzuki, *Transport. Res. Rec. J. Transport. Res. Board* **1855** (2003) 49.
28. Y. Wang, M. Papageorgiou and A. Messmer, *Transport. Res. C-Emerg. Technol.* **14** (2006) 190.
29. S. Fan, M. Herty and B. Seibold, *Netw. Heterogeneous Media* **9** (2014) 239.
30. S. Blandin, J. Argote, A. M. Bayen and D. B. Work, *Transport. Res. B-Methodological* **52** (2013) 31.
31. F. van Wageningen-Kessels, H. van Lint, K. Vuik and S. Hoogendoorn, *EURO J. Transport. Logis.* **4** (2014) 445.
32. A. Sumalee, R. X. Zhong, T. L. Pan and W. Y. Szeto, *Transport. Res. B-Methodological* **45** (2011) 507.
33. S. E. Jabari, J. Zheng and H. X. Liu, *Transport. Res. B-Methodological* **68** (2014) 205.
34. C. Wang and Z. Ye, *J. Intel. Transport. Syst.* **20** (2016) 428.
35. M. V. D. Voort, M. Dougherty and S. Watson, *Transport. Res. C-Emerg. Technol.* **4** (1996) 307.
36. K. Eunjeong, A. Jinyoung and K. E. Yi, *Sensors* **16** (2016) 147.
37. L. Zhang, J. Ma and C. Zhu, *J. Inform. Comput. Sci.* **9** (2012) 5101.
38. J. Guo, W. Huang and B. M. Williams, *Transport. Res. C-Emerg. Technol.* **43** (2014) 50.
39. M. J. Deng and S. R. Qu, *Comput. Intel. Neurosci.* **7** (2015) 1.
40. X. Jiang and H. Adeli, *J. Transport. Eng.* **131** (2020) 771.
41. E. I. Vlahogianni, M. G. Karlaftis and J. C. Golias, *Transport. Res. C-Emerg. Technol.* **13** (2005) 211.
42. W. Zheng and D. H. Lee, *J. Transport. Eng.* **132** (2006) 114.
43. K. Y. Chan, T. S. Dillon, J. Singh and E. Chang, *IEEE Trans. Intel. Transport. Syst.* **13** (2012) 644.
44. A. Cheng, X. Jiang, Y. Li, Cheng, C. Zhang and H. Zhu, *Phys. A: Stat. Mech. Appl.* **466** (2017) 422.
45. H. H. Xie, X. H. Dai and Y. Qi, *J. Traffic Transport. Eng.* **14** (2014) 87.
46. Z. Zheng and D. Su, *Transport. Res. C-Emerg. Technol.* **43** (2014) 143.
47. B. Frank, *IEEE Trans. Intel. Transport. Syst.* **16** (2015) 3393.
48. J. Wang, W. Deng and Y. Guo, *Transport. Res. C-Emerg. Technol.* **43** (2014) 79.
49. J. J. Tang, H. Wang, Y. H. Wang, X. Y. Liu and F. Liu, *Transport. Res. Rec.* **2443** (2014) 21.

50. Y. Gu, D. Wei and M. Zhao, *Lect. Notes Electric. Eng.* **113** (2012) 59.
51. Q. Shang, C. Y. Lin, Z. S. Yang, Q. C. Bing and X. Y. Zhou, *Plos One* **11** (2016) 0161259.
52. C. Quek, M. Pasquier and B. B. Lim, *IEEE Trans. Intel. Transport. Syst.* **7** (2006) 133.
53. L. Li, S. He, J. Zhang and B. Ran, *J. Adv. Transport.* **50** (2017) 2029.
54. W. B. Zhang, Q. Yong, K. Henrickson, J. J. Tang and Y. H. Wang, *Transportmetrica* **13** (2017) 467.
55. F. Guo, R. Krishnan and J. Polak, *J. Intel. Transport. Syst.* **21** (2017) 214.
56. C. Wang and Z. Ye, *J. Intel. Transport. Syst.* **20** (2015) 428.
57. F. G. Habtemichael and M. Cetin, *Transport. Res. C-Emerg. Technol.* **66** (2016) 61.
58. X. Q. Wang, C. F. Shao, C. Y. Yin, X. Ji and L. Guan, *J. Beijing Jiaotong Univ.* **42** (2018) 79.