Transportation Research Part C 62 (2016) 21–34

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

A spatiotemporal correlative *k*-nearest neighbor model for short-term traffic multistep forecasting

Pinlong Cai [a,b], Yunpeng Wang [a,b], Guangquan Lu [a,b,*], Peng Chen [a,b], Chuan Ding [a,b], Jianping Sun [c]

[a] Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, School of Transportation Science and Engineering, Beihang University, Beijing 100191, China

[b] Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, SiPaiLou #2, Nanjing 210096, China

[c] Beijing Transportation Research Center, Beijing 100073, China

ARTICLE INFO

Article history:
Received 8 May 2015
Received in revised form 27 September 2015
Accepted 7 November 2015
Available online 7 December 2015

Keywords:
Short-term traffic forecasting
k-nearest neighbor model
Spatiotemporal correlation
Gaussian weighted Euclidean distance

ABSTRACT

The *k*-nearest neighbor (KNN) model is an effective statistical model applied in short-term traffic forecasting that can provide reliable data to guide travelers. This study proposes an improved KNN model to enhance forecasting accuracy based on spatiotemporal correlation and to achieve multistep forecasting. The physical distances among road segments are replaced with equivalent distances, which are defined by the static and dynamic data collected from real road networks. The traffic state of a road segment is described by a spatiotemporal state matrix instead of only a time series as in the original KNN model. The nearest neighbors are selected according to the Gaussian weighted Euclidean distance, which adjusts the influences of time and space factors on spatiotemporal state matrices. The forecasting accuracies of the improved KNN and of four other models are compared, and experimental results indicate that the improved KNN model is more appropriate for short-term traffic multistep forecasting than the other models are. This study also discusses the application of the improved KNN model in a time-varying traffic state.

1. Introduction

Real-time traffic data can be obtained with the development of intelligent traffic systems (García-Ortiz et al., 1995). However, lag time is detected when such data are applied to formulate strategies for managing and controlling traffic (Smith et al., 2002). Therefore, short-term traffic conditions must be forecasted according to available real-time data. Short-term traffic forecasting has become an interesting research topic in this field. Administrators can manage traffic networks and effectively ensure normal operation with the aid of reliable forecasting data. Travelers can also decide on departure time or travel routes easily.

Several statistical models have been applied extensively in short-term traffic forecasting, including the time series model (Williams and Hoel, 2003; Lee and Fambro, 2007), Kalman filter model (Guo et al., 2014), nonparametric regression method (Smith and Demetsky, 1997; Smith et al., 2002; Zheng and Su, 2014), Bayesian model (Wang et al., 2014), support vector machine regression model (Zhang and Chen, 2010; Hu et al., 2011; Wang and Shi, 2013), and hidden Markov model (Qi and Ishak, 2014). Many artificial intelligence algorithms have also been used effectively to this end. The most typical model established is the neural network model (Smith and Demetsky, 1994; Xie and Zhang, 2006; Dong et al., 2010; Hou, 2011;

* Corresponding author at: Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, School of Transportation Science and Engineering, Beihang University, Beijing 100191, China.

http://dx.doi.org/10.1016/j.trc.2015.11.002

Abdi et al., 2012; Leng et al., 2013; Ma et al., 2015a). Besides, Huang and Sadek (2009) proposed a spinning network model that is similar to human memory. Zhang et al. (2014a) developed a hierarchical fuzzy rule-based system that is optimized by genetic algorithms.

Considerable research has concentrated on comparing an alternative forecasting model with other models given the ready availability of different types of data and the unique features of various models (Vlahogianni et al., 2014). Artificial intelligence algorithms can overcome several problems (e.g., data failure) and provide black box solutions. Model performance constrains the quality of trained data, and these models are difficult to extend from one application to another (Van and Van, 2012). By contrast, the traditional statistical models are simple to implement and are applicable to numerous road segments; however, these models hardly forecast accurately when used alone. The nonparametric regression method has more portability, higher accuracy, and a simpler structure than the parametric models do (Smith and Demetsky, 1997; Zheng and Su, 2014); as a data-driven approach, the former is also suitable for short-term traffic forecasting in urban road networks because this method adapts to the complexity of traffic signals through flexible restructuring (Vlahogianni et al., 2014). An example of the characteristic nonparametric regression method is the k-nearest neighbor (KNN) model, which is easy to implement because the process of training data and estimating parameters is simple. Nonetheless, the search algorithm and method of forecasting result integration in this model should be improved.

The majority of previous studies conducted single-step forecasting depending on the limited data regarding a single road. The duration of such forecasting is less than 15 min, which is generally relatively short to help travelers complete one trip on the currently complex road networks. Several scholars have recently attempted to investigate multistep forecasting; however, the performance of such forecasting deteriorates rapidly with an increase in the number of steps when traditional forecasting methods are employed. Thus, researchers analyzed the relationship among road segments by considering much spatiotemporal data collected from several road segments in a road network. Min and Wynter (2011) considered the distance and average speed of the links in reflecting the spatial characteristics of a road network; these indicators remained accurate for up to 1 h in 12 time steps. Considering the influence of multiple links, Sun et al. (2012) proposed the Bayesian classical model based on the Gaussian regression process for short-term traffic forecasting in urban road networks. Zhang et al. (2014b) established a neural network model of radial basic function by analyzing the traffic flow relationship between a specific road segment and other road segments. Haworth and Cheng (2012) predicted travel time in the central London section by combining the nuclear regression model with the KNN model. Kamarianakis et al. (2012) enhanced the classic time series model and examined a spatiotemporal correlation by analyzing traffic flow variables and nonlinear dynamics. Zou et al. (2014) merged spatial and temporal travel time information to predict travel time within 1 h. The aforementioned researchers considered the spatiotemporal data of nearby road segments for short-term traffic forecasting; however, these scholars were unable to quantize the spatiotemporal correlation among road segments clearly in their forecasting models. Thus, an improved KNN model is proposed in the present study based on the spatiotemporal correlation of road segments.

The present study applies a new criterion of equivalent distances to redefine the contact among road segments and uses spatiotemporal state matrices to identify traffic states. Then, the proposed model enhances the computations of nearest neighbor distance and the integration of forecasting results through the Gaussian weighted method to overcome the defect of the original KNN model. Peak and off-peak times are set within one day given the dynamic flows in a road network. In addition, a time-varying model is reasonably used to forecast short-term traffic states involving different traffic characteristics in various periods. Finally, a deviation compensation method is introduced to adjust the forecasting result further. The maximum forecasting time in the study is 1 h (12 time steps).

This paper is organized as follows: Section 2 proposes the improved KNN model based on spatiotemporal correlations. The equivalent distance, spatiotemporal state matrix, and Gaussian weighted methods are also introduced in this section. Section 3 determines the parameters in the model and compares the performance of the improved KNN model with that of other models with the real data collected. Section 4 discusses the application of the improved KNN model in a time-varying traffic state and the deviation adjustment of the forecasting results. Section 5 concludes the study and proposes directions for future research.

2. Methodology

2.1. Original KNN model

The KNN model for short-term traffic forecasting aims to identify the current state of the traffic network and integrates generations of similar historical states as forecasting results. The specific steps of the original KNN model are described as follows.

First, historical data are used to build the sample database. Either traffic flow or vehicle speed is usually selected as the critical parameter. An appropriate vector space is also defined to describe the current and historical traffic states. Then, the Euclidean distances between all sample and current data are calculated to generate the k-sample data, the distances among which are regarded as the KNNs. Finally, future traffic states are forecasted by averaging generations of KNNs (Smith and Demetsky, 1997).

The original KNN model has been applied in many studies, and several analogous weaknesses should be improved upon. The results of the original KNN model are usually hysteretic in time series and lack prediction accuracy because this model utilizes oversimplified methods and does not consider data from nearby road segments. The current study considers the spatiotemporal correlation of road networks and adopts the Gaussian weighted method to enhance the original KNN model.

2.2. Improved KNN model considering spatiotemporal correlation

The specific steps in the improved KNN model proposed in this study are listed as follows.

First, historical speed data are used to build the sample database. Second, the equivalent distances between the forecasted road segment and other road segments are calculated to determine the spatiotemporal correlative road segments, whose time vectors constitute the spatiotemporal state matrices used to describe traffic states. Third, the equivalent distances between all sample data and current data are calculated according to the Gaussian weighted Euclidean distance to select the KNNs. Finally, the Gaussian weighted average method is applied to obtain the final forecasting results by integrating generations of KNNs.

2.2.1. Equivalent distances and the spatiotemporal state matrix

The traffic states of road segments in a road network tend to be influenced by their upstream and downstream movements. For example, congestions are often initiated at one or more road segments and spread to other road segments after a period of time, thereby resulting in regional congestion (Ma et al., 2015b). The characteristic evolution of the early peak congestion in Liuliqiao District is shown in Fig. 1; the red[1], yellow, and green lines represent the congested, slightly congested, and smooth road segments, respectively. The majority of previous researchers considered only the data from a single road segment in short-term traffic forecasting; by contrast, this study incorporates the spatiotemporal correlation among road segments and uses much data from nearby road segments in a road network region.

The spatiotemporal correlation is determined by analyzing the structure of a road network and the characteristics among the time series of real data. A physical road network is divided into several road segments to collect floating car speed data. On the basis of this division, a connective hierarchy of road segments is established, and the schematic of this hierarchy is depicted in Fig. 2. The first grade ($g = 1$) represents the forecasted road segment; the second grade ($g = 2$) includes the road segments directly connected to the forecasted road segment; and the third grade ($g = 3$) corresponds to the road segments that are directly connected to the second grade. The remaining grades ($g = 4, 5, 6, \ldots$) are similarly described. The other grades may include several road segments, with the exception for the first grade.

The well-known physical distance concept represents the physical property of a road network. However, this concept does not reflect the traffic flow property among road segments. The correlation coefficient is another common concept that is calculated according to the time series of road segments and can reflect the traffic flow property. Connective grade is an artificial concept that reflects the spatial correlations among road segments. This study defines equivalent distance as a criterion that is used to determine correlations between the forecasted road segment and other related road segments. Thus, this distance is a composite variable with physical and data properties; it is denoted as *dist* and is expressed as follows:

$$dist = \begin{cases} (h \cdot g)^{1-r} & g \neq 1 \\ 1 & g = 1 \end{cases} \tag{1}$$

where h is the physical distance among road segments in a road network. This study considers the midpoints of segments to ensure convenient calculation; g is the connective grade of a road segment and r is the correlation coefficient between the historical time series of two road segments that are determined using real data. If X and Y represent two road segments and their time series are $\{x_1, x_2, \ldots, x_N\}$ and $\{y_1, y_2, \ldots, y_N\}$, respectively, then the equation is written as Eq. (2), where N is the length of the time series.

$$r = \frac{Cov(X,Y)}{\sqrt{DX} \cdot \sqrt{DY}} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{2}$$

The value of *dist* is the measure of the spatiotemporal correlation between the forecasted road segment and other related road segments. For the forecasted road segment, $g = 1$, $h = 0$, and $r = 1$. Thus, $dist = 1$ is the minimum of *dist*. This value increases with an increase in that of h, g, or $(1 - r)$.

The general calculation procedure is as follows: first, a suitable maximum value of g is selected to generate a connective road network containing a sufficient number of road segments. Second, the correlation coefficients of time series are calculated between the forecasted road segment and all the road segments in a connective road network. Third, the equivalent distances are calculated with Eq. (1). Finally, a suitable threshold value of equivalent distance is computed to select related road segments from the connective road network. The chosen road segments are combined into a road segment series following the order of equivalent distances. The number of selected road segments increases with an increase in the threshold value, which depends on specific circumstances in the road network. This process is detailed in Section 3.2.

On the basis of the one-dimensional state vector time series established in the original KNN model, this study establishes a two-dimensional spatiotemporal state matrix $V(m, n)$ by using a time dimension and a spatial dimension to determine traf-

[1] For interpretation of color in Fig. 1, the reader is referred to the web version of this article.

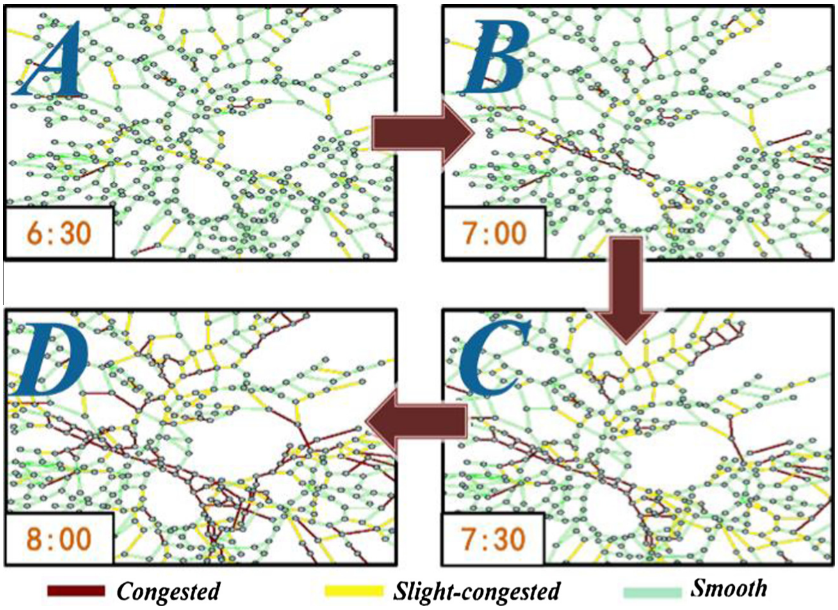


Fig. 1. Congestion spread in a road network.

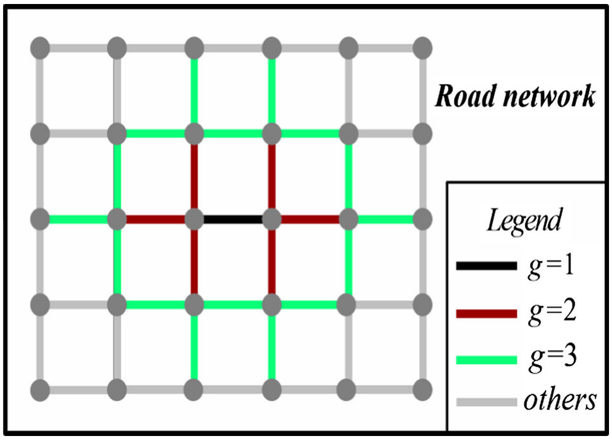


Fig. 2. Schematic of the hierarchy in a road network.

fic state. m is the length of the time series, and n is the number of road segment series (including the forecasted road segment). The elements of the state matrix represent the speed of each road segment in certain time steps. The basic form of $V(m, n)$ is presented in Eq. (3). The traffic state of each road segment at a certain time step is determined by utilizing the data of m time steps (including $m-1$ previous time steps) and n-related road segments (including $n-1$ nearby road segments).

$$V(m,n) = \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{m,1} & v_{m,2} & \cdots & v_{m,n} \end{bmatrix} \tag{3}$$

2.2.2. Gaussian weighted Euclidean distance

KNN models usually use Euclidean distances to calculate the distance of traffic parameters between the current state and all the historical states to identify historical states that are similar to those of the nearest neighbors for forecasting. However, the changing trends of traffic state cannot be reflected by Euclidean distances alone. Zheng and Su (2014) proposed the use of correlation coefficient distance to select the nearest correlated neighbors; this approach focuses on the changing trends of traffic state and disregards the absolute distance between two states. Thus, this study proposes a new method, i.e., Gaussian weighted Euclidean distance, to determine the similarities between two spatiotemporal state matrices whose weights are set on the basis of a Gaussian function.

The time-weighted matrix is defined as W_t, the space-weighted vector is defined as W_s, and their elements are defined as w_t and w_s, respectively. Their forms are shown in Eqs. (4) and (5). Let V and $V_p(m,n)$ represent the current state and the pth historical state, respectively; the similarity degree (SD) between these states is determined by calculating the Gaussian weighted Euclidian distance (Eq. (6)).

$$W_t = \begin{pmatrix} w_{t,1} & & & \\ & w_{t,2} & & \\ & & \ddots & \\ & & & w_{t,m} \end{pmatrix} \quad \text{where } w_{t,i} = \frac{1}{4\pi a_1^2}\exp\left(-\frac{|t_i - t_m|^2}{4a_1^2}\right), \quad i \in [1,m] \tag{4}$$

$$W_s = \begin{pmatrix} w_{s,1} & & & \\ & w_{s,2} & & \\ & & \ddots & \\ & & & w_{s,n} \end{pmatrix} \quad \text{where } w_{s,i} = \frac{1}{4\pi a_2^2}\exp\left(-\frac{|dist_j|^2}{4a_2^2}\right), \quad j \in [1,n] \tag{5}$$

$$SD_p = \|W_t V W_s - W_t V_p W_s\|_2 = \|W_t(V - V_p)W_s\|_2 \tag{6}$$

where t_m is the current time step, t_i represents the $(m-i)$ time steps before t_m, and $dist_j$ is the equivalent distance between the jth road segment and the forecasted road segment. The values of time-weighted parameter a_1 and the space-weighted parameter a_2 are optimized according to real data.

Fig. 3 displays the performance levels of the different proposed distance algorithms. The mean absolute percentage errors (MAPEs) of the improved KNN model applying Gaussian weighted Euclidean distance are less than those of the other two distance algorithms in most forecasted road segments. Thus, the Gaussian weighted Euclidean distance is effectively utilized in the proposed model.

2.2.3. Adjustment of the forecasting results

The final step in the KNN model involves selecting the KNNs of the current state and forecasting through the integration of the KNN generations. In the traditional KNN models, the mean values of generations are usually calculated as the forecasting results. However, the use of weighted mean values is more accurate than that of mean values because of the different SDs of the nearest neighbors. This study applies the Gaussian function to set the weight for each of the nearest neighbors. The weights must be normalized, and the weight of the qth nearest neighbor is defined as λ_q. The relevant equation is expressed as follows:

$$\lambda_q = \frac{1}{4\pi a_3^2}\exp\left(-\frac{|SD_q|^2}{4a_3^2}\right) \tag{7}$$

where a_3 is the result adjustment weighted parameter, which is determined with real data. Thus, the final forecast result in the $(t_m + l)$th time step is defined as F_{t_m+l}. The relevant equation is written as follows:

$$F_{t_m+l} = \left(\sum_{s=1}^{k} \hat{v}_{q,t_{m+l}} \cdot \lambda_q\right) \Bigg/ \sum_{s=1}^{k} \lambda_q, \quad q \in [1,k] \tag{8}$$

where $\hat{v}_{q,t_{m+l}}$ is the forecast speed of the qth nearest neighbor in several time steps following the current time step t_m.

3. Case study

3.1. Data preparation

3.1.1. Data source

The data used to evaluate the performance of the proposed model were the floating car speed data collected in Liuliqiao District, Beijing (Fig. 4). The road network was divided into 1004 road segments, including expressways, arterials, and other roads. The data were collected over four weeks in 2013 (i.e., two weeks in March and two weeks in September), and the time interval was 5 min. Each recorded piece of floating car speed data included record time, vehicle speed (space mean speed), and road segment length. This study selected a representative area in a district with 30 road segments, and the effective data were derived over 20 weekdays from 00:00 to 23:55 (288 time steps in 1 day). Thus, 20 days' worth of data were employed in this study, and these data are divided into three parts. In the first part, 10 days are allocated to build the historical database. The second part covers five days to determine the parameters and adjust the methods. The third part allots five days to evaluate the models.

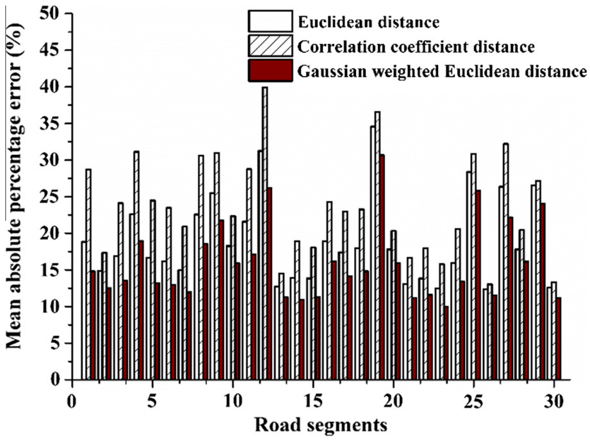


Fig. 3. Performance levels of the proposed model given different distance algorithms.

3.1.2. Data processing

The performance of short-term traffic forecasting models often relies on complete data; however, data were missing and abnormal in the original set. Thus, these data should be processed before forecasting (Haworth and Cheng, 2012; Li et al., 2013). The present study applied two methods to address the missing data for each road segment. First, these missing data were replaced with data from the same period as collected on other days. Second, linear interpolation calculation was performed with adjacent time steps if data remained missing. The abnormal data for certain road segments, excluding those in the 95% confidence interval range, were identified from all speeds at the same time of day.

The design speed in a road network was varied for different types of road segments. The proposed model in this study aims to forecast short-term traffic state involving all types of road segments in a road network; thus, the data were normalized with the following equation:

$$\hat{v}_{i,j} = v_{i,j}/fs_j, \quad i \in [1, m_0],\ j \in [1, n_0] \tag{9}$$

where m_0 is the maximum number of time steps, n_0 is the sum of the road segments, $\hat{v}_{i,j}$ is the normalized speed of the jth road segment in the ith time step, $v_{i,j}$ is the actual speed, and fs_j is the free-flow speed of the jth road segment, which is constant at all times.

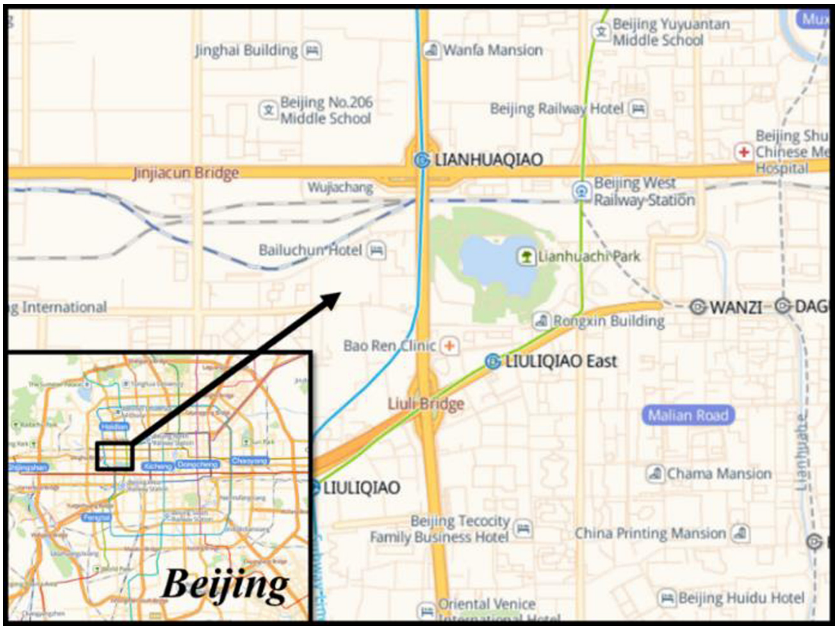


Fig. 4. Liuliqiao District.

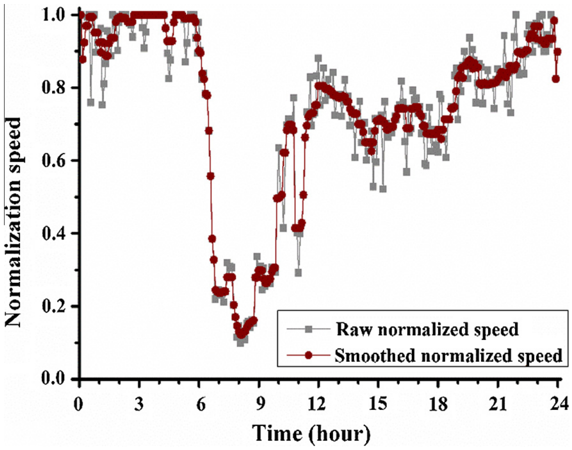


Fig. 5. Normalized speeds of the road segment (ID 25744).

The free-flow speed value is usually defined via two indicators, namely, the speed limit of the road at free-flow speed and the 85th percentile speed value. In the latter, 85% of the speeds in a certain road are below a certain value. This method was adopted in this study to utilize the massive amounts of real data. This study assumed that the normalized range increases from 0 to 1; thus, the normalized speeds greater than 1 were assigned a value of 1. Finally, a median filter with seven time steps was applied to smooth the data. Fig. 5 shows the comparison of raw and smoothened normalized speeds and indicates that the latter fluctuates less than the former. The smoothened normalized speeds also reflect the continuous change in traffic flow more accurately than raw normalized speeds do.

3.2. Parameter calibration

Model performance can be evaluated by calculating their MAPE and root mean square error (RMSE) values; the relevant formulas are shown in Eqs. (10) and (11). MAPE, which is the most important index, reflects the relative errors of the models, whereas RMSE reflects the fact that significant errors in certain time steps may aggravate problems that should be addressed (Odeck, 2013). We also use the median of the mean absolute percentage error (MDAPE) in different road segments to avoid generating several road segments with abnormal data.

$$\mathrm{MAPE} = \frac{1}{M \cdot L \cdot Num} \sum_{\alpha=1}^{M} \sum_{l=1}^{L} \sum_{cnt=1}^{Num} \frac{|F^{\alpha}_{t_m+l,cnt} - \hat{v}^{\alpha}_{t_m+l,cnt}|}{\hat{v}^{\alpha}_{t_m+l,cnt}} \tag{10}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{M \cdot L \cdot Num} \sum\nolimits_{\alpha=1}^{M} \sum\nolimits_{l=1}^{L} \sum\nolimits_{cnt=1}^{Num} \left| F^{\alpha}_{t_m+l,cnt} - \hat{v}^{\alpha}_{t_m+l,cnt} \right|^2} \tag{11}$$

where M is the number of forecasted road segments ($M = 30$), L is the maximum number of forecast time steps ($L = 12$), l is the forecast time step ($l \in [1, L]$, 5 min between two continuous time steps), $cnt \in [1, 2, \ldots, Num]$ denotes the number of data in each road segment, and $F^{\alpha}_{t_m+l,cnt}$ and $\hat{v}^{\alpha}_{t_m+l,cnt}$ are the forecasting and real values of the αth road segment, respectively, at the lth time step after current time step t_m.

The performance of the forecasting model is evaluated with MAPE in parameter calibration. The data used in this calibration belong to the second part, including 30 road segments in 5 days. The time steps m in the spatiotemporal state (see Eq. (1)) are set to 12 (1 h). The number of correlative road segments n is determined according to the threshold value of *dist*. A time constraint must also be set to select the nearest neighbors; this constraint enhances accuracy and computing speed (Zheng and Su, 2014). A time constraint of 3 h (36 time steps) was implemented in the current study; for example, if the spatiotemporal state includes the time steps from 07:00 to 07:55 (12 time steps in total), then 20 probable neighbors are obtained from 06:00 to 08:45 (36 time steps in total) in a historical day.

According to the preliminary experiment results, the maximum value of g (Eq. (1)) is set as 3. Fig. 6 indicates that the performance of the forecasting model changes with an increase in the threshold value of *dist*. The MAPEs decrease rapidly when the threshold value ranges between 1.0 and 3.5 but decrease gradually when this value exceeds 3.5. Thus, the threshold value is determined to be 3.5 considering the variability of the MAPEs and computing speed. Table 1 shows the numerical test results for the road segment (ID 25744). 10 correlative road segments are studied in this road segment. The *dist* values are less than 3.5.

Fig. 7(a) depicts the relationships between forecasting accuracy and the time-weighted parameter a_1 (Eq. (4)). MAPE increases with an increase in a_1, and performance is optimized when the latter is sufficiently small. Thus, a_1 is determined to be 0.01. According to Eq. (4), $w_{t,m}$ is larger than the other time weights if a_1 is small enough; this finding indicates that the

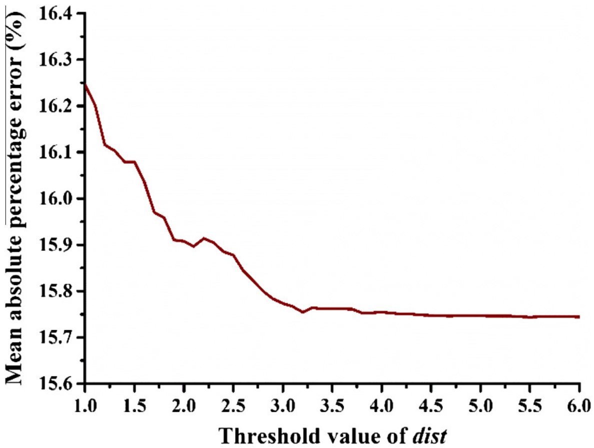


Fig. 6. Change in MAPEs with an increase in the threshold value of *dist*.

Table 1
Numerical test results for *dist* (ID 25744).

ID of road segment	*g*	*h*	*r*	*dist*	ID of road segment	*g*	*h*	*r*	*dist*
25744	1	0.00	1.00	1.00	14552	3	289.64	0.91	1.88
14091	2	205.76	0.84	2.62	14554	3	240.19	0.78	4.33
25741	2	251.64	0.98	1.14	25228	3	361.86	0.86	2.75
25742	2	216.73	0.92	1.64	25641	3	808.65	0.90	2.10
25743	2	317.43	0.98	1.15	25728	3	294.25	0.78	4.59
14399	3	490.59	0.97	1.21	25731	3	235.96	0.77	4.66
14455	3	494.40	0.98	1.18					

data regarding previous time steps in the spatiotemporal state matrix do not influence short-term traffic forecasting, i.e., the current time step is sufficient. Fig. 7(b) exhibits the relationships between forecasting accuracy and the space-weighted parameter a_2 (Eq. (5)). When a_2 increases, MAPEs decrease rapidly and then increase gradually with the fluctuations. The a_2 value is set as 1.01 in the model; moreover, the data on the road segment series in the current time step effectively reflect the current traffic state through the parameter values. Fig. 7(c) shows the relationship between the variability of the result adjustment parameter a_3 (Eq. (7)) and the forecasting accuracy of the proposed model with real data. Model accuracy is optimized when $a_3 = 0.49$.

The number of KNNs (Eq. (8)) has been determined for most existing KNN models; this information significantly affects forecasting accuracy. The present study integrated the generations of nearest neighbors, and effective data can be obtained when k value increases. The actual results are shown in Fig. 7(d). MAPEs decrease rapidly when k is less than 40 and increase gradually when k is greater than 40. Thus, the appropriate k value is determined to be 40. Table 2 presents the calibration results of all parameters in the models.

3.3. Performance comparison

In this study, the proposed model is compared with the historical average model, Elman neural network (Elman-NN), least squares support vector machine (LS-SVM), and original KNN model. The historical average model is the earliest model applied to traffic forecasting in the past century; Elman-NN and LS-SVM are extensively applied in short-term traffic forecasting applications (Dong et al., 2010; Hou, 2011; Hu et al., 2011; Zhang and Chen, 2010); and the original KNN model has effectively been implemented by many researchers, such as Smith et al. (2002) and Zheng and Su (2014).

Fig. 8 shows that MAPE values change with different forecasting steps in the five models. The improved KNN model performs better than the other models in each step, and the MAPE of the proposed model is approximately 2.96% greater than that of the original KNN model. This result indicates that the improved KNN model based on spatiotemporal correlation is effective for short-term traffic multistep forecasting. Tables 3 and 4 present the comparison of the numerical results (including MAPE, RMSE, and MDAPE) for single-step and multistep forecasting.

4. Discussion

The improved KNN model is detailed in this section, including the application of the space state vector and the enhancement of model performance with time-varying parameters and with the deviation compensation method.

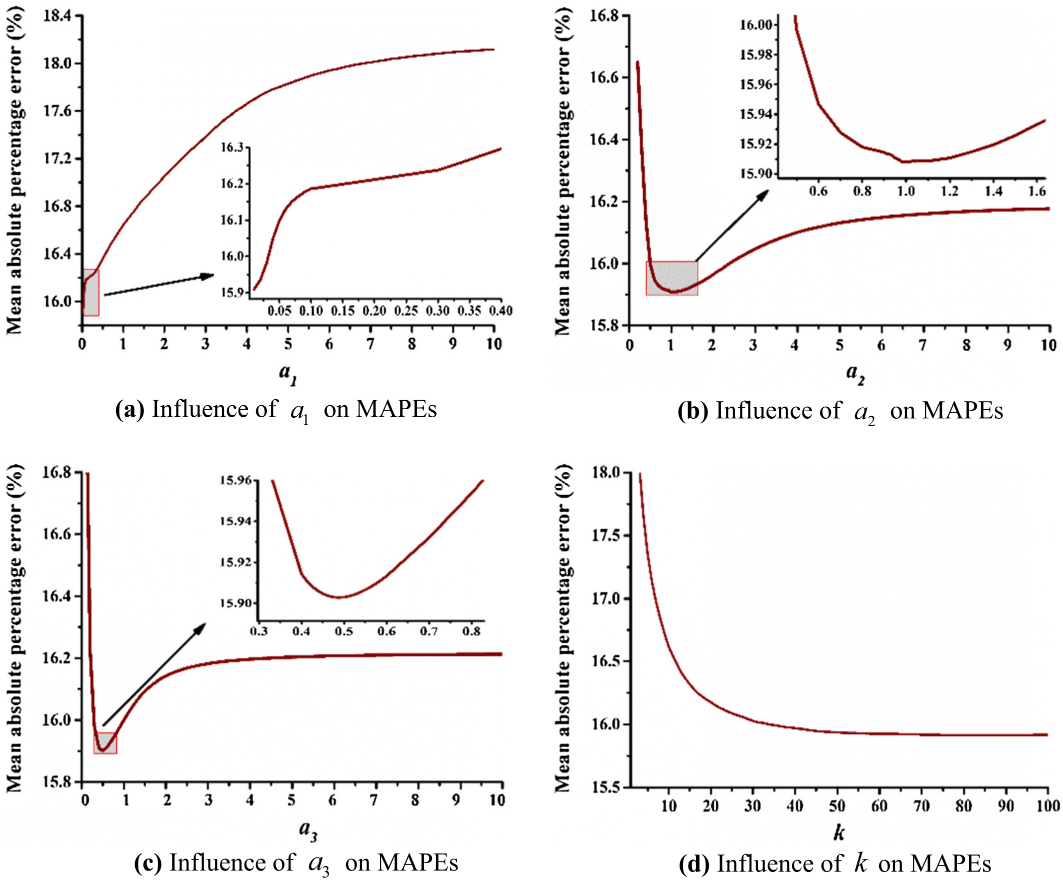


Fig. 7. Influence of the parameters on the improved KNN models.

Table 2
Calibration results of the model parameters.

Parameters	Eq.	Values	Description
a_1	(4)	0.01	Time-weighted parameter
a_2	(5)	1.01	Space-weighted parameter
a_3	(7)	0.49	Result adjustment weighted parameter
k	(8)	40	Number of nearest neighbors

4.1. Space state vector

The improved KNN model considers the time series of the road segment to be forecasted and the influence of related road segments. Thus, this model performs well in terms of multistep short-term traffic forecasting. The spatiotemporal correlation in the improved KNN model is reflected in several aspects, including in the use of equivalent distances to redefine the relationship between the road segment and the spatiotemporal state matrix. In the process, the traffic state can be described accurately. The value of the time-weighted parameter a_1, which affects the time series in the state matrix, is small; therefore, the time series spatiotemporal state matrix includes the current step and the spatiotemporal state matrix degenerates into a one-dimensional space state vector. The traffic state in the road segment to be forecasted can be determined according to space state vector, including the data and the nearby related road segments in the nearest time step.

4.2. Time-varying parameters

Traffic state characteristics change continuously in a real traffic network, and Fig. 9(a) depicts the mean lines of 30 road segments in 10 days (288 time steps in 1 day). Traffic states can be categorized into three types. The normalized speeds at night are almost higher than those observed during the day; moreover, peak and off-peak times are common in daytime

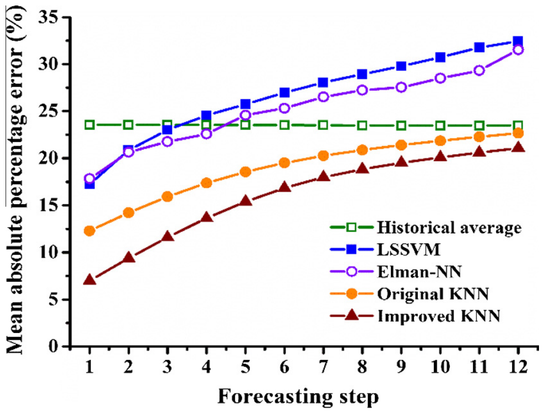


Fig. 8. Performance comparison among different models.

Table 3
Comparison of the numerical results for single-step forecasting.

Model	MAPE (%)	RMSE	MDAPE (%)
Historical average	23.57	0.15	22.41
LS-SVM	17.25	0.11	16.73
Elman-NN	17.83	0.11	16.47
Original KNN	12.28	0.08	11.35
Improved KNN	6.99	0.05	6.38

Table 4
Comparison of the numerical results for multistep forecasting.

Model	MAPE (%)	RMSE	MDAPE (%)
Historical average	22.39	0.15	22.39
LS-SVM	26.68	0.17	26.34
Elman-NN	25.28	0.16	25.52
Original KNN	18.93	0.12	17.53
Improved KNN	15.99	0.10	14.19

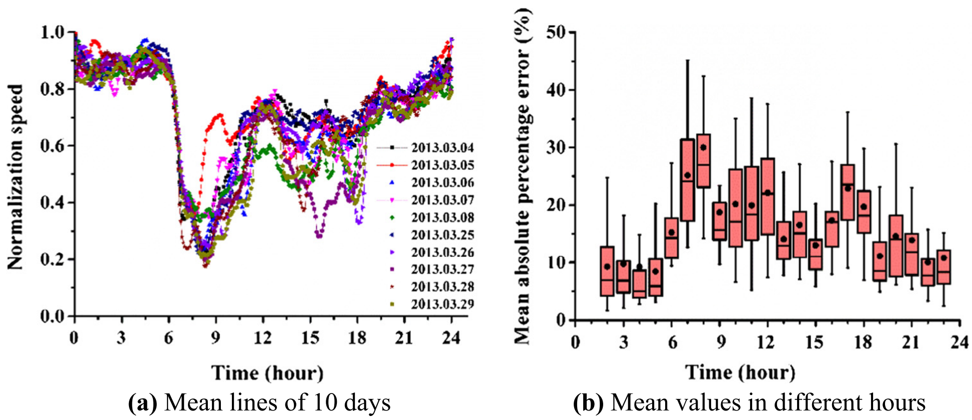


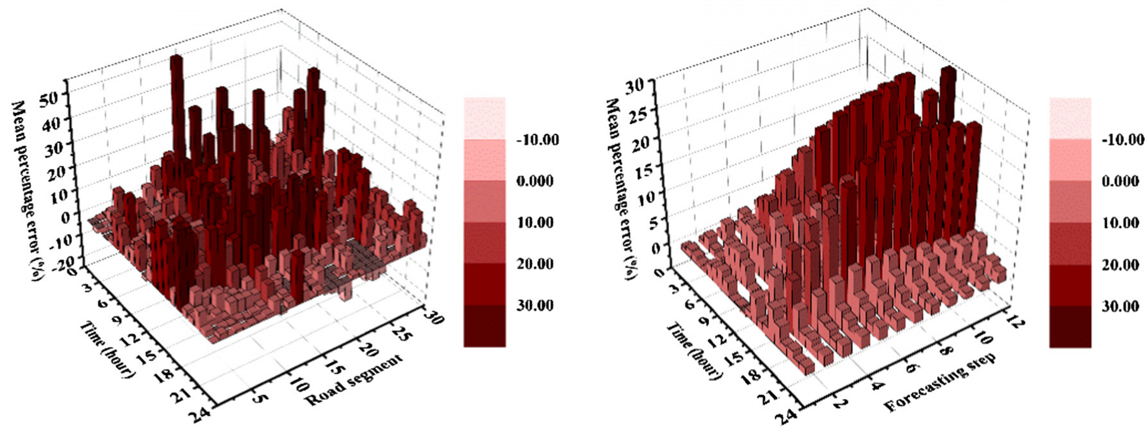
Fig. 9. Time-varying characteristics.

because of the varying travel behavior in different periods. Congestions are unavoidable during peak hours, and the variances in speed during these times are always greater than those during off-peak times. As a result, traffic states are difficult to forecast.

Fig. 9(b) displays the forecasting accuracy of 30 road segments in different hours (each day has 22 h and excludes the first and the last hours of a day). Forecast accuracy is more volatile during peak hours than during non-peak hours. Furthermore,

Table 5
Determination of parameters in different classes.

Class	a_1	a_2	a_3	k
1	0.01	0.75	0.35	40
2	0.01	0.53	0.54	40
3	0.01	0.44	0.54	40



(a) Influences of different road segments **(b)** Influences of different forecasting steps

Fig. 10. Influences of various factors on MPEs.

the forecasting deviations vary with time, and the same is true for the optimal parameters in the improved KNN model. A significant linear negative correlation was calculated between the mean normalized speed of each hour and MAPEs as per Pearson's correlation analysis; the correlation coefficient is -0.747. Given the complex traffic state in the peak period, the traffic state is difficult to forecast.

This study conducts a hierarchical cluster analysis to divide 22 h (12 time steps in 1 h) into 3 classes; thus, different traffic state features can be distinguished in 1 day. The first class contains {8th and 9th}, which represents the early peak time that often reports severe congestions. The second class contains {10th, 11th, 15th, 16th, and 18th}, and the third class contains {2nd, 3rd, 4th, 5th, 6th, 7th, 12th, 13th, 14th, 17th, 19th, 20th, 21st, 22nd, and 23rd}.

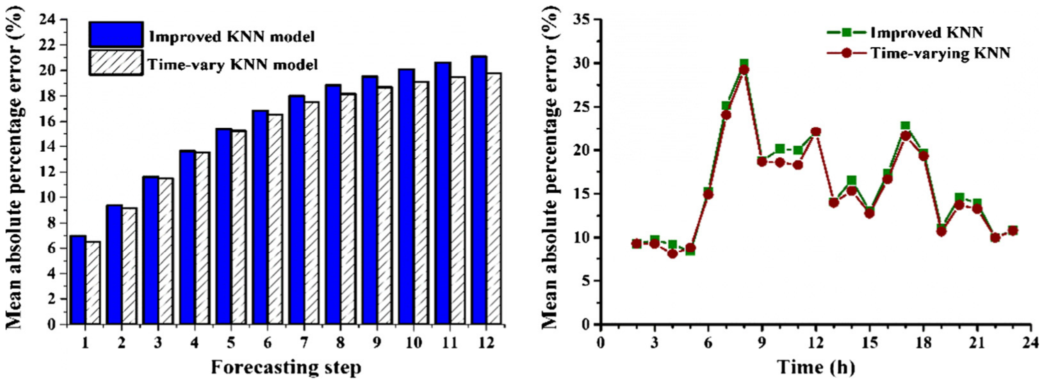
The values of the model parameters (i.e., a_1, a_2, a_3, and k in Eqs. (4)–(8)) should vary with different classes. The determination process is similar to that presented in Section 3.2, and the results are shown in Table 5. The values of a_1 and k do not vary with time, whereas a_2 and a_3 values differ in dissimilar classes. In the peak hours, the value of a_2 is larger than those of the other parameters. This outcome indicates that the nearby road segments influence the objective road segments strongly. When a_3 increases, the nearest neighbors contribute more to the final forecasting result than before.

4.3. Deviation compensation

Fig. 10 shows the influence of different road segments and forecasting steps on the mean percentage error (MPE). The relevant formula is expressed as Eq. (12), and its parameters are similar to those in Eqs. (10) and (11). MPEs increase rapidly with an increase in forecasting steps in several periods; however, this pattern is not universal across all road segments. Theoretically, MPEs are often small because the negative and positive values are typically offset. By contrast, anomalous MPEs indicate systematic underestimation or overestimation in the forecasting process (Odeck, 2013). To reduce these occurrences, the current study compensates for the forecasting results by averaging the historical MPEs of each road segment. These MPEs are recorded chronologically along with each forecasting step; the records cover 268 forecasting time steps a day.

$$\mathrm{MPE} = \frac{1}{M \cdot L \cdot Num} \sum_{\alpha=1}^{M} \sum_{l=1}^{L} \sum_{cnt=1}^{Num} \frac{\left(F^{\alpha}_{t_m+l,cnt} - \hat{v}^{\alpha}_{t_m+l,cnt}\right)}{\hat{v}^{\alpha}_{t_m+l,cnt}} \tag{12}$$

The improved KNN model can be enhanced further using the time-varying methods (i.e., time-varying parameters and the deviation compensation method). The new model is called the time-varying KNN model. The performance levels of the improved KNN model and of the time-varying KNN model are compared, as shown in Fig. 11. Fig. 11(a) indicates that both



(a) Comparison of each forecasting step **(b)** Comparison for each hour

Fig. 11. Performance comparison between the improved and the time-varying KNN models.

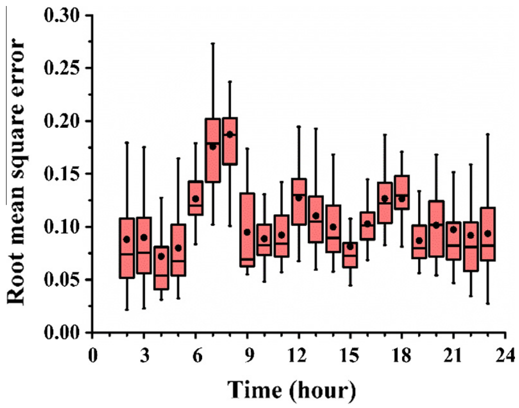


Fig. 12. RMSEs in different hours.

models perform similarly during the first few forecasting steps; however, the time-varying KNN model becomes more accurate with the progression of the forecasting steps. Fig. 11(b) also shows that the time-varying KNN model typically performs better than the improved KNN model; however, the improvement of the time-varying KNN model on the improved KNN model is insignificant. The MAPEs of all road segments decrease by 0.57% when the time-varying methods are used.

The forecasts made during peak time are unsatisfactory because of the complex traffic state in this period, although these predictions improve on previous ones (Fig. 11(b)). In addition, we evaluated the forecast error with MAPE; however, this approach may generate a misleading result when the real value is small. A low normalized speed increases MAPE during the peak hours. Meanwhile, the error gap between the peak and non-peak hours narrows if we assess the forecast error according to another criterion, such as RMSE (Fig. 12).

5. Conclusions

Many scholars have studied short-term traffic forecasting based on much traffic data (i.e., speed and traffic flow), and numerous models have been proposed. The KNN model is one such common statistical model. This study aims to improve the forecasting accuracy of KNN models in multistep forecasting and to enhance model performance in time-varying traffic conditions.

This work improves the original KNN model based on spatiotemporal correlation and discusses the details of its application to actual road networks. The improved KNN model follows these processes: first, the proposed model uses equivalent distances to define contacts among road segments. These segments contain the actual distances, link relations, and coefficients among road segments. Equivalent distance is dynamic and changes with the traffic data collected. Second, the model employs a spatiotemporal state matrix to identify the traffic state in the road network instead of the one-dimensional time series utilized in previous research. This identification reflects the spatiotemporal correlation of the road network. Third, the Gaussian function is applied to set the weights for the model several times for optimizing forecasting performance.

This study evaluates the effectiveness of the proposed model by using the floating car speed data collected from an actual road network. MAPE and RMSE are introduced to compare the performance of the improved KNN with that of other models.

The result indicates that the proposed improved methods are effective and that the improved KNN model performs better than the other models do.

This work also discusses the use of the improved KNN model under the time-varying traffic condition. The forecasting result shows that the time-varying KNN model established with the time-varying parameter and deviation compensation methods is more adaptive to actual application in short-term traffic forecasting than other models are.

The main objective of this research is to present a new KNN method that considers the spatiotemporal correlation for traffic state forecasting. However, this study has several limitations. Equivalent distance is a critical factor in the analysis of the spatiotemporal correlation between road segments; therefore, a more appropriate form of equivalent distance than that shown in Eq. (1) should be developed through further experimentation. The forecast accuracy during the peak time should be improved as well.

In future works, the spatiotemporal correlative KNN model can be expanded to forecast more traffic parameters using more types of traffic data. In contrast to existing research (Lei et al., 2014; Park and Haghani, 2015), the evolution of the forecasting process for road network state or travel time can be studied by considering the spatiotemporal correlation, which may provide intuitive data for before-trip and on-trip travelers.

Acknowledgments

The authors would like to thank the Beijing Transportation Research Center for providing the data used in this study. This study was supported by the National Basic Research Program of China (Project 2012CB725404).

References

Abdi, J., Moshiri, B., Abdulhai, B., Sedigh, A.K., 2012. Forecasting of short-term traffic-flow based on improved neurofuzzy models via emotional temporal difference learning algorithm. Eng. Appl. Artif. Intell. 25 (5), 1022–1042.
Dong, C., Shao, C., Xiong, Z., et al, 2010. Short-term traffic flow forecasting of road network based on Elman neural network. J. Transport. Syst. Eng. Inform. Technol. 10 (1), 145–151.
García-Ortiz, A., Amin, S.M., Wotton, J.R., 1995. Intelligent transportation systems—enabling technologies. Math. Comput. Modell. 22 (4–7), 11–81.
Guo, J., Huang, W., Williams, B.M., 2014. Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification. Transport. Res. Part C: Emerg. Technol. 43, 50–64.
Haworth, J., Cheng, T., 2012. Non-parametric regression for space–time forecasting under missing data. Comput. Environ. Urban Syst. 36 (6), 538–550.
Hou, J., 2011. Prediction for network traffic based on Elman neural network. Comput. Simul. 28 (7), 154–157.
Hu, Y., Wu, C., Liu, H., 2011. Prediction of passenger flow on the highway based on the least square support vector machine. Transp. Res. J. Vilnius Gediminas Tech. Univ. Lithuanian Acad. Sci. 26 (2), 197–203.
Huang, S., Sadek, A.W., 2009. A novel forecasting approach inspired by human memory: the example of short-term traffic volume forecasting. Transport. Res. Part C: Emerg. Technol. 17 (5), 510–525.
Kamarianakis, Y., Shen, W., Wynter, L., 2012. Real-time road traffic forecasting using regime-switching space–time models and adaptive LASSO. Appl. Stoch. Models Bus. Ind. 138, 446–451.
Lee, S., Fambro, D.B., 2007. Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting. Transport. Res. Rec. 1678 (1999), 179–188.
Lei, F., Wang, Y., Lu, G., et al, 2014. A travel time reliability model of urban expressways with varying levels of service. Transport. Res. Part C: Emerg. Technol. 48, 453–467.
Leng, Z., Gao, J., Zhang, B, Liu, X., Ma, Z., 2013. Short-term traffic flow forecasting model of optimized BP neural network based on genetic algorithm. In: 2013 32nd Chinese Control Conference (CCC).
Li, L., Li, Y., Li, Z., 2013. Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. Transport. Res. Part C: Emerg. Technol. 34, 108–120.
Ma, X., Tao, Z., Wang, Y., et al, 2015a. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transport. Res. Part C: Emerg. Technol. 54, 187–197.
Ma, X., Yu, H., Wang, Y., et al, 2015b. Large-scale transportation network congestion evolution prediction using deep learning theory. PLoS ONE 10 (3), e0119044.
Min, W., Wynter, L., 2011. Real-time road traffic prediction with spatio-temporal correlations. Transport. Res. Part C: Emerg. Technol. 19 (4), 606–616.
Odeck, J., 2013. How accurate are national road traffic growth-rate forecasts?—the case of Norway. Transp. Policy 27, 102–111.
Park, H., Haghani, A., 2015. Optimal number and location of bluetooth sensors considering stochastic travel time prediction. Transport. Res. Part C: Emerg. Technol. 55, 203–216.
Qi, Y., Ishak, S., 2014. A hidden Markov model for short term prediction of traffic conditions on freeways. Transport. Res. Part C: Emerg. Technol. 43, 95–111.
Smith, B.L., Demetsky, M.J., 1994. Short-term traffic flow prediction: neural network approach. Transport. Res. Rec. 1453, 98–104.
Smith, B.L., Demetsky, M.J., 1997. Forecasting freeway traffic flow for intelligent transportation systems application. Transport. Res. Part A: Policy Pract. 31 (1), 61.
Smith, B.L., Williams, B.M., Oswald, R.K., 2002. Comparison of parametric and nonparametric models for traffic flow forecasting. Transport. Res. Part C: Emerg. Technol. 10 (4), 303–321.
Sun, S., Huang, R., Gao, Y., 2012. Network-scale traffic modeling and forecasting with graphical lasso and neural networks. J. Transport. Eng. 138 (11), 1358–1367.
Van, L., Van, H., 2012. Short term traffic and travel time prediction models, in artificial intelligence applications to critical transportation issues. Transport. Res. E – Circ.
Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C., 2014. Short-term traffic forecasting: where we are and where we're going. Transport. Res. Part C: Emerg. Technol. 43, 3–19.
Wang, J., Deng, W., Guo, Y., 2014. New Bayesian combination method for short-term traffic flow forecasting. Transport. Res. Part C: Emerg. Technol. 43, 79–94.
Wang, J., Shi, Q., 2013. Short-term traffic speed forecasting hybrid model based on Chaos-Wavelet analysis-support vector machine theory. Transport. Res. Part C: Emerg. Technol. 27, 219–232.
Williams, B.M., Hoel, L.A., 2003. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: theoretical basis and empirical results. J. Transport. Eng. 129 (6), 664–672.
Xie, Y., Zhang, Y., 2006. A wavelet network model for short-term traffic volume forecasting. J. Intell. Transport. Syst. 10 (3), 141–150.
Zhang, C., Chen, L., 2010. Traffic flow combining forecast model based on least squares support vector machine. J. Hunan Inst. Eng. 20 (4), 56–58.

Zhang, L., Jia, Y., Niu, Z., Li, H., 2014a. Research on short-term traffic flow forecasting for junction of isomerism road network based on dynamic correlation. Procedia – Soc. Behav. Sci. 138, 446–451.

Zhang, X., Onieva, E., Perallos, A., et al, 2014b. Hierarchical fuzzy rule-based system optimized with genetic algorithms for short term traffic congestion prediction. Transport. Res. Part C: Emerg. Technol. 43, 127–142.

Zheng, Z., Su, D., 2014. Short-term traffic volume forecasting: a *k*-nearest neighbor approach enhanced by constrained linearly sewing principle component algorithm. Transport. Res. Part C: Emerg. Technol. 43 (1), 143–157.

Zou, Y., Zhu, X., Zhang, Y., et al, 2014. A space–time diurnal method for short-term freeway travel time prediction. Transport. Res. Part C: Emerg. Technol. 43, 33–49.