k-Nearest Neighbor Model for Multiple-Time-Step Prediction of Short-Term Traffic Condition

Bin Yu[1]; Xiaolin Song, Ph.D.[2]; Feng Guan, Ph.D.[3]; Zhiming Yang, Ph.D.[4]; and Baozhen Yao[5]

Abstract: One of the most critical functions of an intelligent transportation system (ITS) is to provide accurate and real-time prediction of traffic condition. This paper develops a short-term traffic condition prediction model based on the *k*-nearest neighbor algorithm. In the prediction model, the time-varying and continuous characteristic of traffic flow is considered, and the multi-time-step prediction model is proposed based on the single-time-step model. To test the accuracy of the proposed multi-time-step prediction model, GPS data of taxis in Foshan city, China, are used. The results show that the multi-time-step prediction model with spatial-temporal parameters provides a good performance compared with the support vector machine (SVM) model, artificial neural network (ANN) model, real-time-data model, and history-data model. The results also appear to indicate that the proposed *k*-nearest neighbor model is an effective approach in predicting the short-term traffic condition. **DOI: 10.1061/(ASCE)TE.1943-5436.0000816.**

Author keywords: Short-term traffic condition; Multi-time-step prediction model; *k*-nearest neighbor; Spatial-temporal parameters.

Introduction

With the acceleration of urbanization and mobilization in China, the number of motor vehicles has risen dramatically, resulting in worse urban traffic conditions and more serious environmental problems. Traffic congestion has become a bottleneck, restricting the development of urban economy. Therefore, it is imminent to alleviate the urban traffic congestion problem. In recent years, the intelligent transportation system (ITS), whose core subsystems are the traffic signal control system (TSCS) and the traffic flow guidance system (TFGS), is applied widely in China. The TSCS manages the road traffic flow state through sign control, whereas the TFGS provides traffic information for the driver to adjust the driving route to achieve traffic flow equilibrium. Proving accurate traffic flow information is the goal of these applications. Therefore, effective short-term traffic-flow prediction is the key to traffic induction and control, which directly affects the performance of the transportation system. This is why an increasing amount of attention among traffic engineers and researchers has been given to the short-term traffic forecasting.

Until the present, a series of technical methods have been proposed to predict the development of short-term traffic flow. These methods can be grouped into two categories: one that employs mathematical or statistical forecasting algorithms, and one that uses modern scientific techniques, each attempting to pursue the strict sense of the mathematical derivation and clear physical sense, but relying on real traffic-flow phenomena fitting effect. Because of the complexity of the problem, it is difficult for the models to express the complex relationship among internal elements. Therefore, many researchers have generally adopted the second category of methods. Zheng et al. (2006) combined the neural network model with the Bayesian model to predict short-term traffic flow. Tan et al. (2009) used three models—moving average model, exponential smoothing model, and autoregressive moving average (MA) — to forecast the time series, and afterward used the neural network to integrate the three prediction results. Yao et al. (2014a) proposed a support vector machine to predict the freeway traffic flow and acquired freeway incident detection. Boto et al. (2010) and Jiang and Adeli (2005) combined wavelet and neural to establish a prediction model. Vlahogianni et al. (2007) addressed the problem of integrating predictive information from multiple locations. A modular neural prediction consisting of temporal genetically optimized structures of multilayer perceptions was developed. To solve the problem of the traffic-flow prediction usually being too complex, Zhong et al. (2012) proposed a traffic-flow-prediction algorithm based on historical frequent pattern. To predict the short-term passenger flow, Wei and Chen (2012) built a model that combined empirical mode decomposition and back-propagation neural networks. Lin et al. (2013) established short-term forecasting of traffic-volume evaluation models, offering a solution to choosing the appropriate prediction method on the basis of the statistical characteristics of the data set. Xu et al. (2015) presented an algorithm based on compressive sensing to estimate the traffic states of the road traffic.

Among these prediction methods, the nonparametric regression model, which can deal with the uncertainties that happen in traffic flow, is a practical one with high prediction accuracy. The *k*-nearest neighbor (*k*-NN) model is a widespread method of nonparametric regression prediction. Many scholars have successfully applied the *k*-NN model in short-term traffic-flow prediction. Zuo et al. (2008) developed a new method to improve the weigh setting in the *k*-NN model. To prevent the interference of discontinuous data in the database, Akbari et al. (2011) established a clustered *k*-NN model. Turochy (2006) combined the condition-monitoring method with

[1]Professor, Transportation Management College, Dalian Maritime Univ., Dalian 116026, P.R.China.

[2]Transportation Management College, Dalian Maritime Univ., Dalian 116026, P.R. China.

[3]Transportation Management College, Dalian Maritime Univ., Dalian 116026, P.R. China.

[4]Transportation Management College, Dalian Maritime Univ., Dalian 116026, P.R. China.

[5]Associate Professor, School of Automotive Engineering, Dalian Univ. of Technology, Dalian 116024, P.R. China (corresponding author). E-mail: yaobaozhen@hotmail.com

Note. This manuscript was submitted on July 25, 2014; approved on September 22, 2015; published online on February 16, 2016. Discussion period open until July 16, 2016; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Transportation Engineering***, © ASCE, ISSN 0733-947X.

the k-NN model, so that the atypical traffic condition could be considered and the forecasting accuracy could be improved. Smith and Demetsky (1997) developed four models—historical average, time series, neural network, and nonparametric regression models—to test the freeway traffic forecasting, and the result showed the nonparametric model outperformed the other models. In a further study, Smith et al. (2002) concluded that the nonparametric regression coupled with heuristic forecast is better than the naïve forecasting technique. William et al. (2006) compared the nonparametric regression (NPR) model with the Gaussian maximum likelihood (GML) model. Based on historical data, the prediction result showed that the NPR model performs better than GML model. On the basis of the k-NN nonparametric regression (k-NN-NPR) model, Chang et al. (2012) built a dynamic multiple-interval traffic-flow prediction model. Their model could make full use of the data collected from advanced data management systems, and recognize the rich information of historical data. Yoon and Chang (2014) compared the k-NN nonparametric regression forecasting methodology with Kalman filtering and seasonal autoregressive integrated moving average (ARIMA), and concluded that the k-NN-NPR model was clearly superior to the two parametric models in terms of both prediction accuracy and the construction of the directionality of temporal state evolution without a time-delayed response. Meng et al. (2015) proposed a two-stage short-term traffic-flow prediction method based on the balanced binary tree (AVL) and advanced k-nearest neighbor (AKNN) techniques. These studies prove that the k-NN model performs well in the short-term traffic-flow forecasting. Based on successful studies, a k-NN model is also applied in this study for the prediction of short-term traffic conditions.

Early studies on traffic condition prediction were limited primarily to single-time-step prediction (i.e., researchers forecasted the traffic flow on the next one-time interval). In recent years, several researchers have paid attention to multistep temporal prediction. Okutani and Stephanedes (1984) used Kalman filtering theory to predict multistep traffic flow. Coric et al. (2012) pointed out that most of the existing state–estimation algorithms were based on Kalman filtering and its variants. They proposed aggregated measurements with signal reconstruction techniques and used them to correct the state estimation at every time step. Smith and Demetsky (1996) established a multiple-interval freeway-traffic-flow forecasting model to predict traffic flow in 15-min intervals for several hours into the future. Lee and Billings (2003) introduced a direct multistep-ahead forecasting for nonlinear time series. Yao et al. (2014b) used a multi-step-ahead prediction method, based on SVM, for rock displacement surrounding a tunnel. Parlos et al. (2000) used a dynamic neural network to perform multistep-ahead predictions while maintaining acceptable single-interval-ahead prediction accuracy. Yao et al. (2015) proposed a hybrid model, based on the k-nearest neighbours method and the Kalman filter, to predict real-time traffic flow. Kirby et al. (1997) compared the neural network model and ARIMA, and the result showed that the prediction accuracy of the purpose-built pattern-based prediction model is better than the other models. Li and Rose (2011) developed three models to predict travel times, and the results provided insight into the relative merits of the proposed model. Thus, this paper also adopts a multiple-time-step prediction model.

In recent years, almost all research has taken temporal information into consideration when predicting short-term traffic conditions. On the basis of these studies, some researchers have also considered spatial information such as upstream and downstream road-link information. However, few of them have illustrated the effects that upstream and downstream road-link information has on the prediction accuracy. Moreover, some research has demonstrated the prediction accuracy of different prediction models, but for multi-time-step prediction of short-term traffic conditions, further discussion of the comparison of several prediction methods is still essential. This paper seeks to make two contributions to the literature. First, it attempts to develop the models to predict short-term traffic conditions based on k-NN models that consider both the temporal and spatial information. The anxiety and travel times of travelers are predicted to be less if they know the operating condition of the road links. The time-varying and continuous characteristics of traffic flow are considered in the proposed models for predicting short-term traffic conditions. Second, k-NN does not require parametric regression and has good generalizability performance. The performances of k-NN and several prediction methods are assessed and compared for multi-time-step prediction of short-term traffic conditions. The performance comparison of several methods can provide valuable insight for researchers and practitioners alike.

The remainder of the paper is organized as follows. First, the paper describes multi-time-step prediction problems. Then, the theory of the k-NN algorithm is introduced. Third, a single-time-step prediction model of short-term traffic conditions, in which spatial-temporal parameters are considered, is proposed. Fourth, the multi-time-step prediction models are constructed. Last, the computational results and conclusions are discussed.

Problem Description

Single-time-step prediction is used to predict traffic conditions on the target road link of the next one time interval. Multi-time-step prediction, which is the extension of the single-time-step prediction, aims to predict the traffic condition of more than one time interval. For short-term traffic condition prediction, multi-time-step prediction can provide travelers with prediction results of longer time.

Fig. 1 shows an example of a three-time-step prediction method of short-term traffic conditions. Assuming the current moment is 7:00 on the current road link, take 5 min as the time interval, the first-step forecasting (i.e., the traditional single-time-step prediction) can predict the traffic condition at 7:05 on the current road link. Then, the second and third-step predictions would predict the traffic condition at 7:10 and 7:15 on current road link, respectively.

k-Nearest Neighbor Model

The k-nearest neighbor is a common nonparametric regression method. It does not need any parametric or model. It is also one of

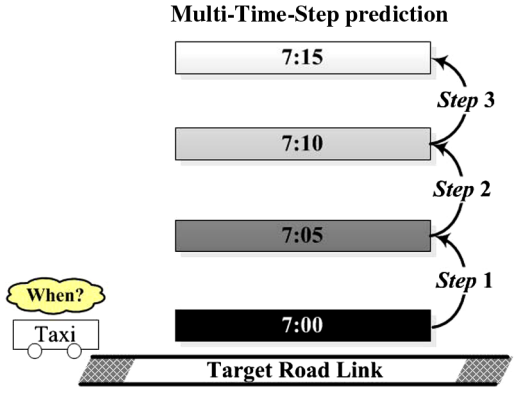


Fig. 1. Multi-time-step prediction based on road link

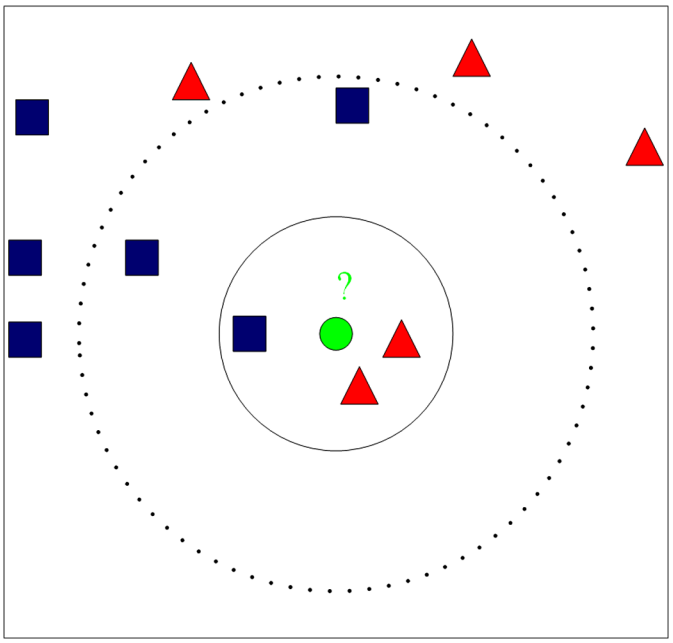


Fig. 2. Example of the k-NN idea

the simplest machine-learning algorithms. The core idea of the method is that if k most similar samples in the feature space belong to some category, then the sample would be determined to belong to this category. Fig. 2 shows an example of the idea of k-NN. There are two kinds of different sample data that are represented by the blue squares and red triangles, respectively. What needs to be determined is the category to which the green circle belongs. It depends on the value of k. If $k = 3$, the three nearest neighbors of the green circle are two red triangles and a blue square. The green circle is determined to belong to the category of the red triangles based on the statistical method. If $k = 5$, the five nearest neighbors of the green circle are two red triangles and three blue squares. Therefore, the green circle belongs to the category of the blue squares based on the statistical method.

Thus, the problem of predicting the short-term traffic condition in this paper is based on the k-NN algorithm. It searches for the matching *nearest neighbors* between historical data and current data according to parameters and data similarity. Then, the searched nearest neighbors are used for prediction. The k-NN algorithm holds that the linkage of data is automatically presented in the database. Therefore, when calculating, only a large volume of data is necessary; no mathematical model and parameters need to be defined in advance. Because no smoothing processing is done to the data, the authenticity of the data is retained. This feature fully embodies the nonparametric characteristic of the short-term traffic condition prediction. The general formula of the k-NN model has been described in detail in the study of Smith et al. (2002) and Akbari et al. (2011). Parameters of the k-NN model include state vector, distance metric, number of nearest neighbor k, and prediction algorithm. In this study, the distance metric and prediction algorithm are preset:

1. The distance metric is used to measure the degree of approximation between the sample data and test data. k data that have the nearest distance between the test data and sample are selected in the k-NN model; these data are then taken as the forecasting data, which will be used in the prediction algorithm. In this study, the Euclidean distance is used as the distance metric. The traffic volume on a road link varies widely from time to time because of the time variability of transport system, so different time periods influence future traffic conditions by a different degree. Thus, different weights should be allocated based on the close degree between time components in the state vector and the forecasting time (i.e., the distance metric of different components should not be same). The longer time gap between the time component and the forecasting time, the smaller the *distance* should be. The correlation coefficient weighting method is used to calculate the distance in this study

$$d_i = \sqrt{\sum_j w_j (V_j - v_{ji})^2} \tag{1}$$

where d_i = distance of group i between the current database and historical database; w_j = weight of state vector j in current database; V_j = value of state vector j in current database; and v_{ji} = value of state vector j in historical database i.

2. The prediction algorithm describes how the searched k groups of nearest neighbors are used to predict the traffic state vector of the next one time period

$$S_m(t+1) = \sum_{g=1}^{k} \frac{d_i^{-1}}{\sum_{g=1}^{k} d_i^{-1}} S_{g,h}(t+1) \tag{2}$$

where $S_m(t+1)$ = average travel speed at time period $t+1$ on road link m; $S_{g,h}(t+1)$ = average speed of nearest neighbor g at time period $t+1$ searched in historical database; and k = number of nearest neighbors.

Prediction Model of Short-Term Traffic Condition

Temporal and Spatial Parameters

Road links do not exist in isolation in urban road networks. The traffic condition on both the upstream and downstream road links may affect the traffic condition of the current road link. Theoretically, considering both the spatial and temporal information is helpful in explaining road information. Most research about short-term traffic condition prediction is focused primarily on studying temporal information; fewer studies consider spatial information as well. These studies include Yu et al. (2010), Kamarianakis and Prastacos (2003), Min and Wynter (2011), and Vlahogianni et al. (2007). According to these successful cases, a k-NN improved model consisting of temporal and spatial parameters is developed in this study. First, the single-time-step prediction models including different state vectors are established in this study, followed by the multi-time-step prediction model.

Single-Time-Step Prediction Model Based on k-NN

Four k-NN models—*temporal* model, *upstream* + *temporal* model, *downstream* + *temporal* model, and *temporal* + *spatial* model—are proposed in this paper to compare the prediction accuracy.

Temporal Prediction Model

Only the temporal information is used as the state vector in the *temporal* model. Fig. 3 shows an example of predicting travel speed in a future time period by using the temporal state vector. Only the temporal state vector on the target road link $S_m(t-n), S_m(t-n+1), \ldots, S_m(t)$ is observed. Then, the travel speed at a future time period $\hat{S}_m(t+1)$ can be predicted.

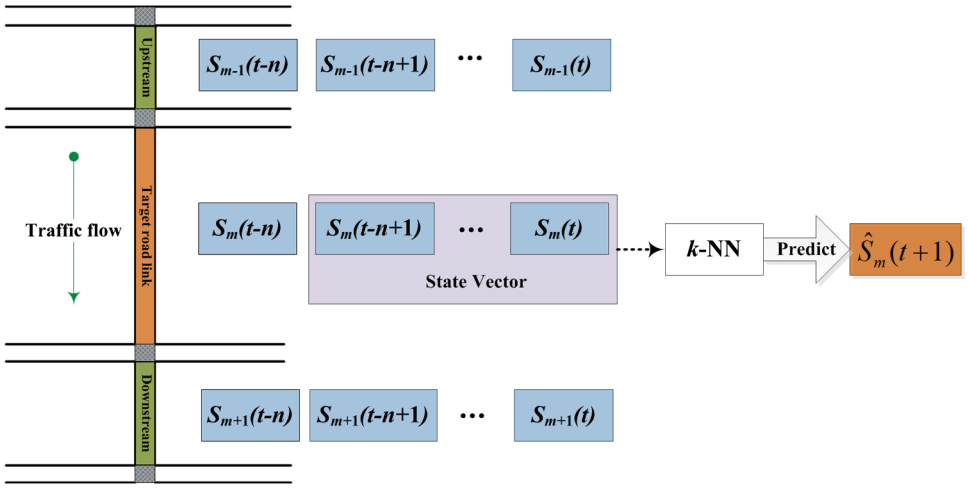


Fig. 3. Temporal prediction model based on k-NN

Upstream + Temporal Prediction Model

In addition to the temporal information of the target road link, the traffic condition on the upstream road link (i.e., road link $m-1$) is also considered in the upstream + temporal model. Fig. 4 illustrates the upstream + temporal model, which includes the traffic condition on both the target and upstream road link.

Downstream + Temporal Prediction Model

In addition to the temporal information of the target road link, the traffic condition on the downstream road link (i.e., road link $m+1$) is also considered in the downstream + temporal model. Fig. 5 illustrates the downstream + temporal model, which includes both the traffic condition on the target and downstream road link.

Spatial + Temporal Prediction Model

In the spatial + temporal model, temporal information on the target road link and the upstream/downstream traffic conditions are considered. As shown in Fig. 6, this model combines both the upstream + temporal model and downstream + temporal model.

Multi-Time-Step Prediction Model Based on Road Link

The multi-time-step prediction model predicts the traffic condition of more than one time interval on each road link (Fig. 7).

The calculation steps are as follows:

1. Step 1: Initialization.
   First, data from the current road link, upstream road link, and downstream road link are gathered.
   Set $t=1$, where t denotes the current time point.
   Set n, where n denotes the maximum number of time steps.
2. Step 2: Establish a single-time-step prediction model.
3. Step 3: Use a single-time-step prediction model to predict the travel speed $\hat{S}_m(t+1)$ at the next time interval $t+1$ on target road link.
4. Step 4: Terminating check of the multi-time-step prediction.
   If exceeding the maximum time ($t >$ n), then stop; otherwise, set $t=t+1$ and go to Step 3, until Step n.

Numerical Study

Data Collection and Processing

In this paper, traffic condition is assessed based on the GPS data of taxis in Foshan, China. Six road links on Foshan Avenue were

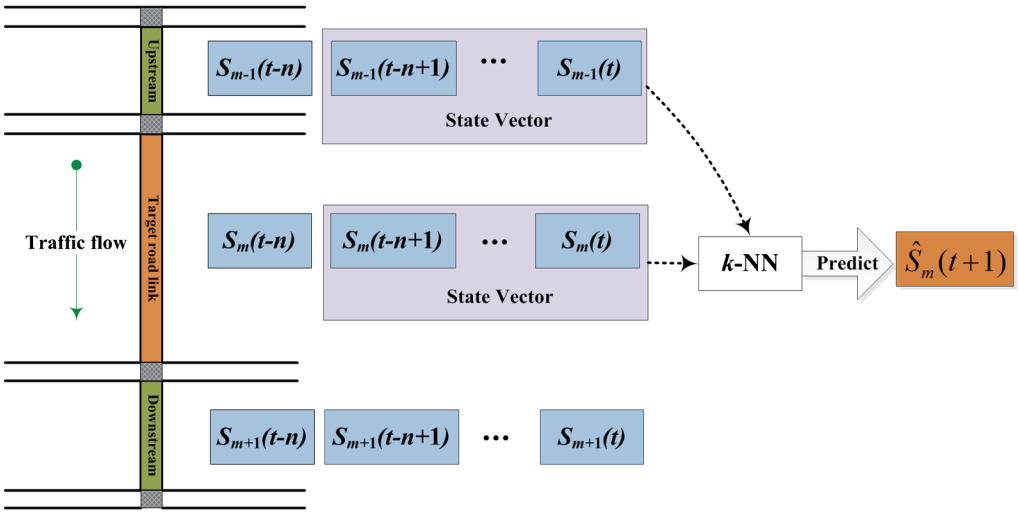


Fig. 4. Upstream+temporal prediction model based on k-NN

J. Transp. Eng., 2016, 142(6): 04016018

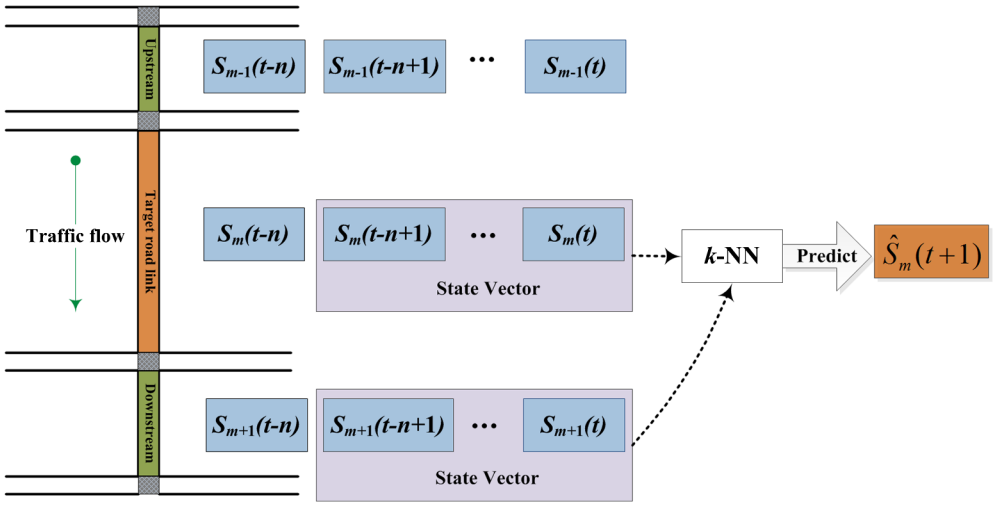


Fig. 5. Downstream + temporal prediction model based on *k*-NN

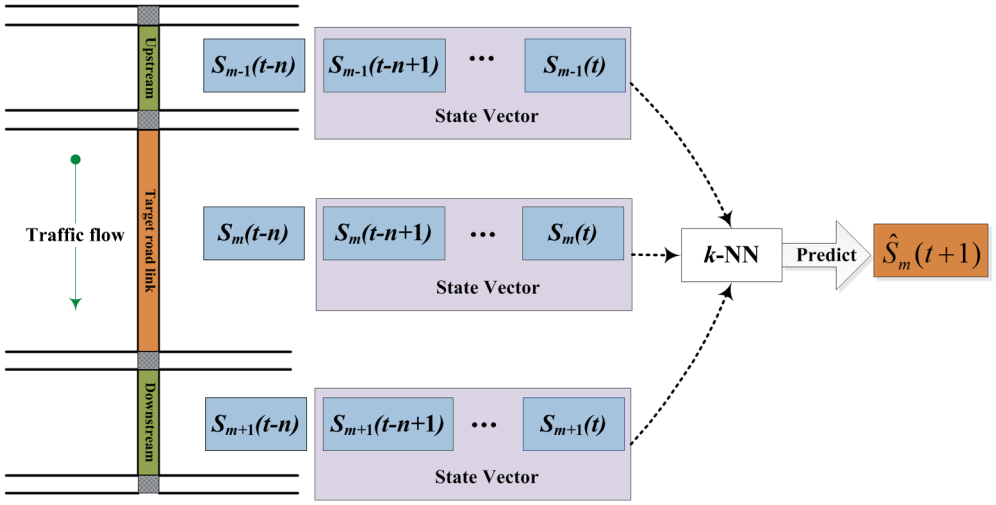


Fig. 6. Spatial + temporal prediction model based on *k*-NN

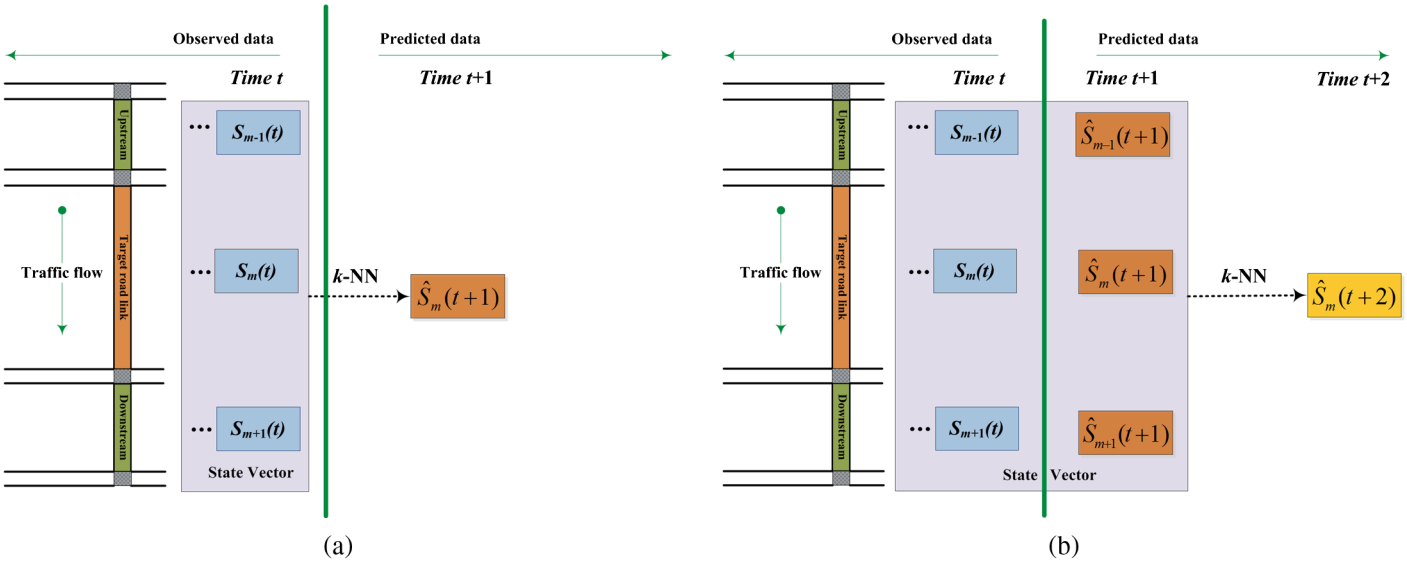


Fig. 7. Multi-time-step prediction model based on road link: (a) first-time-step prediction; (b) second-time-step prediction

J. Transp. Eng., 2016, 142(6): 04016018

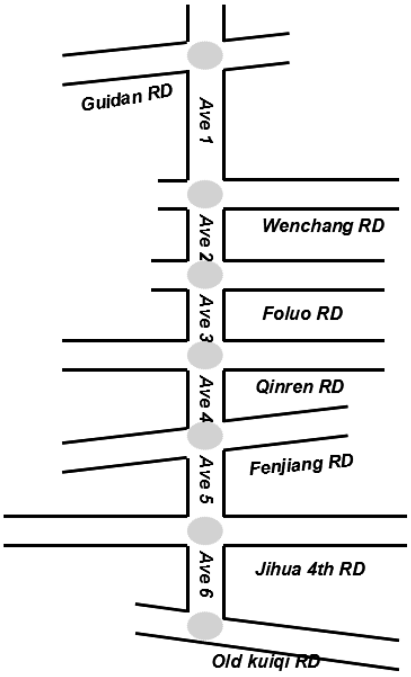


Fig. 8. Spatial location of test links

chosen for the study: Guidan Road–Wenchang Road, Wenchang Road–Foluo Road, Foluo Road–Qinren Road, Qinren Road–Fenjiang Road, Fenjiang Road–Jihua 4th Road, and Jihua 4th Road–Old Kuiqi Road. The spatial location and road information of these six road links are shown in Fig. 8. The total length of the six-road link is 7.6 km. Information from 400 global positioning systems (GPS) was collected during the morning peak (7:00–9:00) on working days from October 8 to November 10, 2012. Fig. 9 illustrates the GPS information of taxis and the matching condition with satellite map.

To provide the basis of the short-term traffic condition prediction, the speed of the taxis in the six road links is needed. The data involved in the taxis' delay caused by picking up and delivering passengers are not needed. This study is limited by not considering the delays caused by the taxis' behavior (i.e., stopping to pick up and deliver happens quite frequently compared with regular passenger cars). However, what the GPS system can supply is a velocity vector with direction. The speed of the taxi at a road link can be computed by the length of the road link that the car traveled across, and the time that the car spent at, the road link

$$S_m(t) = \sum_{c=1}^{p_{m,t}} S_{cm}(t)/p_{m,t} \tag{3}$$

where $S_m(t)$ = average driving speed at the time t on the road link m. Thus, the average driving speed on a road link during each time period could be calculated. Finally, 109,828 valid data were calculated. Table 1 provides the data from each road link, and Fig. 10 shows the average speed of each road link during the different time periods.

Model Determination

Next, the database is divided into three categories: sample set, test set, and forecasting set. For the 1-month data of taxis in Foshan, data from Monday to Wednesday were taken as the sample data, and the data on Thursday and Friday are taken as the test data and forecasting data, respectively. Overall, the test data and forecasting data took up approximately 20% of the sample database, respectively, and the rest were used as the sample set. The mean absolute percentage error (MAPE) was used as the performance measure to evaluate and display the efficacy of the proposed models in this paper. The MAPE was used to assess the difference between the observed and forecasted traffic condition.

Number of Nearest Neighbor k

The prediction accuracy based on the k-NN model is highly contingent on the value of k. The temporal model is considered as the test model to determine the value of $k(1 \le k \le 8)$. The road links used for modeling purposes should have the upstream/downstream, so the MAPEs of Avenues 2, 3, 4, and 5 were calculated using the technique of bootstrapping by 10 times, and the results are shown in Fig. 11.

The MAPEs of these four road links are highest when $k = 1$. This demonstrates that the searched one nearest neighbor cannot comprehensively explain all of test data. When $k = 2$, the MAPEs decline rapidly and the prediction accuracy increases. When k is more than 3, the overall volatility is not big, especially when $k \succ 5$, the volatility is small. Comparing the MAPE with a different k, the forecasting errors for the road link 2 are lowest when $k = 3$, and the forecasting errors for the road links 4 and 5 are lowest when $k = 4$. These results indicate that driving speed information on the road links can be explained well when $k = 4$, and the value of k is set as 4 in this paper.

k-NN Model of Different State Vectors

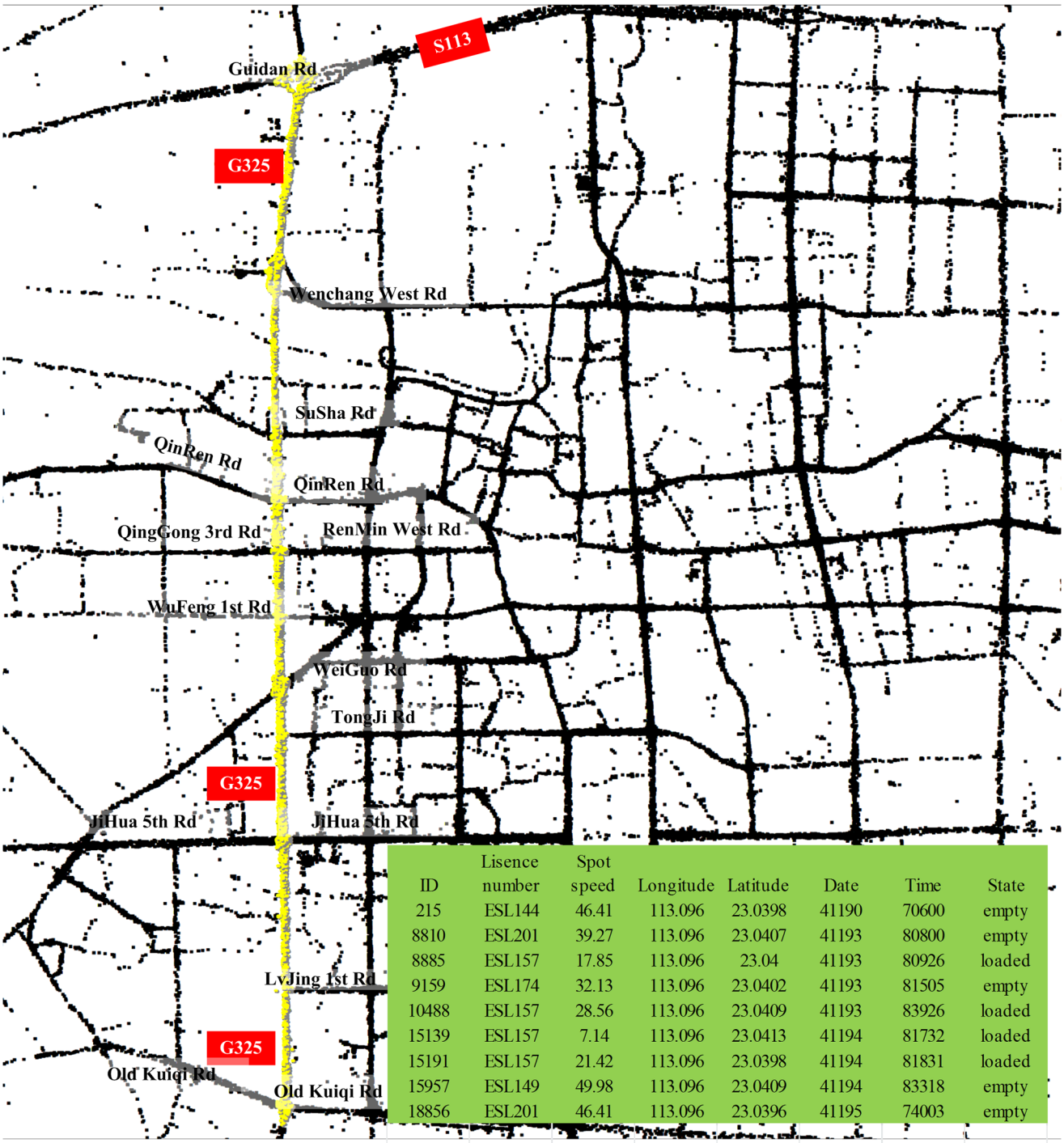
After the computations were complete, the authors found that the traffic flow before several periods could not stand for the traffic condition of the upstream road link. Therefore, n was set as 1, and the data of two periods were used for the prediction, i.e., $S(t)$ and $S(t-1)$. To determine the final k-NN prediction model for the short-term traffic flow, nine models were constructed based on the state vectors mentioned previously (Table 2). All possible models except the models in which only $S_{m-1}(t-1)$ of the two state vectors on the upstream road link were used. Similarly, the authors did not construct the models considering only $S_{m+1}(t-1)$ of the two state vectors on the downstream road link. When comparing the data of the two periods in the prediction, whether for the upstream road link or the downstream road link, $S(t)$ has a greater contribution to prediction accuracy than $S(t-1)$.

The average prediction errors of the different models for four road links are shown in Fig. 12. In addition, to illustrate the advantages of the models, the standard deviations of different models were also calculated.

Fig. 12 shows that Models 4–9 have a better prediction accuracy and stability than Models 1–3, indicating that the downstream road link information could increase the prediction accuracy. Moreover, when the number of state vectors upstream of the road link and current road link is the same, the accuracy of Model 7 is higher than Models 1 and 4, the accuracy of Model 8 is higher than Models 2 and 5, and the accuracy of Model 9 is higher than Models 6 and 3. This illustrates further the importance of downstream road link information on prediction accuracy.

In contrast, when the number of state vectors downstream of the road link and current road link is the same, too much upstream road link interferes with the prediction accuracy. Model 3 is worse than Models 1 and 2, and Model 6 is worse than Models 4 and 5. Compared with Models 7 and 8, Model 9 performed the worst.

Overall, Model 8 is found to be the best one with the least MAPE and standard deviation. This demonstrates that its prediction



ID	Lisence number	Spot speed	Longitude	Latitude	Date	Time	State
215	ESL144	46.41	113.096	23.0398	41190	70600	empty
8810	ESL201	39.27	113.096	23.0407	41193	80800	empty
8885	ESL157	17.85	113.096	23.04	41193	80926	loaded
9159	ESL174	32.13	113.096	23.0402	41193	81505	empty
10488	ESL157	28.56	113.096	23.0409	41193	83926	loaded
15139	ESL157	7.14	113.096	23.0413	41194	81732	loaded
15191	ESL157	21.42	113.096	23.0398	41194	81831	loaded
15957	ESL149	49.98	113.096	23.0409	41194	83318	empty
18856	ESL201	46.41	113.096	23.0396	41195	74003	empty

Fig. 9. GPS information of taxis

accuracy and stability are the best, which is why it is used in this paper.

Results

To evaluate the performance of the *k*-NN model in multi-time-step prediction, the model is used to forecast the traffic conditions on road links for five steps into the future. Eleven time intervals from 7:10 to 8:00 were selected as the sample data. Fig. 13 shows the forecasting errors of the model for the five-time-step prediction on Road Link 4. The average MAPEs of the first step to the fifth step predictions were 10.37, 13.58, 16.81, 19.49, and 20.53%, respectively. The predicted data calculated from the previous steps were used successively as the state vectors in the following steps. Meanwhile, the errors generated from the predicted data gradually accumulated as more predicted data in the state vectors for multi-step prediction models. Therefore, the general trend of MAPE during multi-time-step prediction increased with the forecasting steps.

Table 1. Data from Each Road Link

Link number	Empty		Loaded	
	Number of data	Average speed	Number of data	Average speed
1	6,868	30.35	16,152	26.61
2	578	31.61	1,625	27.91
3	4,301	32.05	10,887	27.8
4	8,058	29.57	25,171	26.8
5	3,978	32.94	12,074	28.84
6	5,406	31.74	14,730	29.21

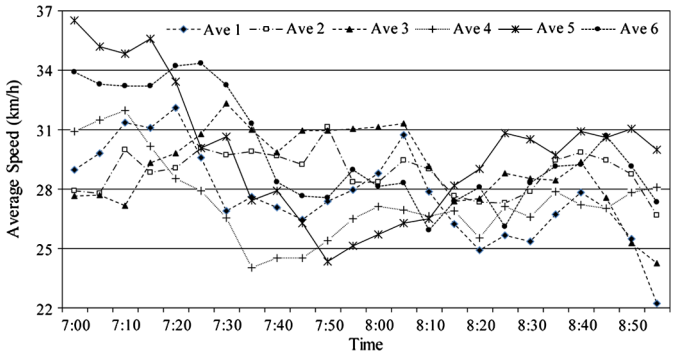


Fig. 10. Average speed of six road links during different periods

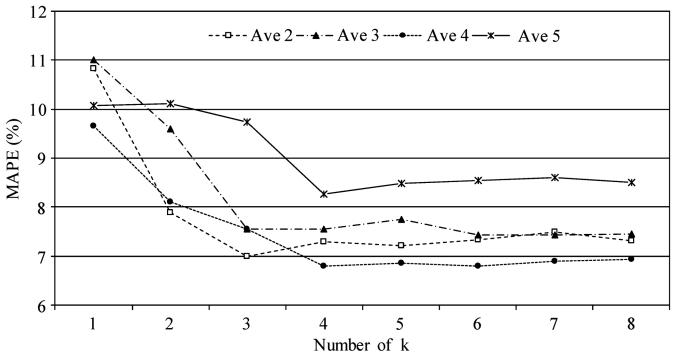


Fig. 11. MAPE of different number of nearest neighbors

Table 2. Models and Corresponding State Vectors

Model	Upstream road link		Target road link		Downstream road link	
	$S_{m-1}(t)$	$S_{m-1}(t-1)$	$S_m(t)$	$S_m(t-1)$	$S_{m+1}(t)$	$S_{m+1}(t-1)$
1	—	—	×	×	—	—
2	×	—	×	×	—	—
3	×	×	×	×	—	—
4	—	—	×	×	×	—
5	×	—	×	×	×	—
6	×	×	×	×	×	—
7	—	—	×	×	×	×
8	×	—	×	×	×	×
9	×	×	×	×	×	×

To compare the accuracy of different step predictions, the five-time-step prediction was executed in each time period on each road link. Then, the average prediction errors of all of the time periods at each step of each road link could be calculated, and the MAPE of

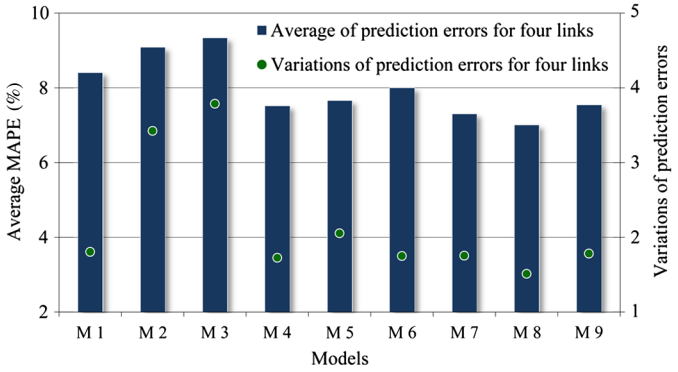


Fig. 12. MAPE of nine models

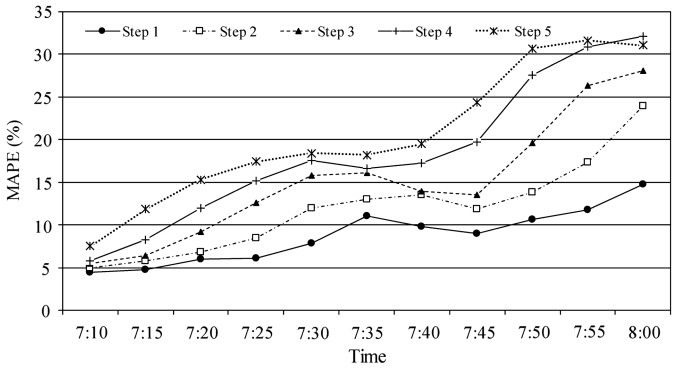


Fig. 13. Multi-time-step prediction on Road Link 4

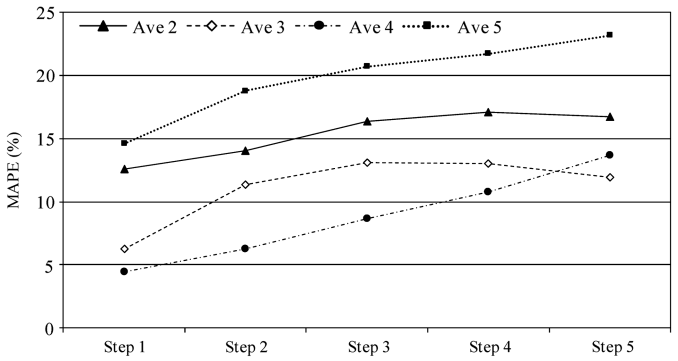


Fig. 14. Multi-time-step prediction of Road Links 2–5

different steps on different road links are shown in Fig. 14. Road links 2–5 showed a similar trend to the Road Link 4 that was calculated previously. The errors of Road Links 4 and 5 increased monotonically; in contrast, Road Links 2 and 3 decreased slightly at the fifth-step prediction. This is because Road Links 2 and 3 are not part of the central business district of Foshan city, so the traffic flow change is less than that of Road Links 4 and 5. Thus, the long-term prediction values tended to flatten, and the errors did not rise noticeably since the fourth step.

Prediction Accuracy Comparison of Different Models

To evaluate the model prediction performance of the multi-time-step prediction model, the five-step k-NN model was compared

with the real-time-data model, history-data model, artificial neural network (ANN) model, and support vector machine (SVM) model.

The formula of the real-time-data model is shown in Eq. (4). n_1 is the number of previous time periods used. The travel speed of time period $t+1$ on the road link m is predicted by the variables of the speed in the previous time period. In this study, n_1 is set as 3. Because three groups of original data are used for the prediction, the final prediction time period starts at 7:25 and ends at 8:55

$$S_m(t+1) = \frac{1}{n_1}[S_m(t) + S_m(t-1) + S_m(t-2) + \cdots S_m(t-n_1+1)] \quad (4)$$

The formula of the history-data model is shown in Eq. (5). $S_m(t_{n_2}+1)$ is the average speed of all taxis during time period n_2; and $S_m(t_i)$ is the average speed of taxis during time interval i. A number of data during a later time period could be used to forecast the travel speed, and in this study's results, the forecasting time period was from 7:15 to 8:45

$$S_m(t_{n_2}+1) = \frac{1}{n_2}\sum_{i=1}^{n_2} S_m(t_i) \quad (5)$$

The input of the ANN model is the same as the one of the proposed k-NN model. After determining the input of the ANN model, a scaled conjugate gradient algorithm was used to train the ANN model. The number of hidden neurons was found to be five in this study. Thus, the final ANN model in this study is the three-layer ANN model and five hidden neurons for travel speed prediction. Based on the the results of the test, the SVM model was found to have a better performance using Model 5 (Table 1), so Model 5 was determined to be the input of the SVM model. The five models were used to predict the traffic conditions of each road link for steps 1–5 into the future, respectively. The average prediction errors of the four road links of each model are shown in Fig. 15.

Fig. 15 shows that the average MAPEs of the history data are generally bigger than those of the real-time-data model. The history-data model has the worst prediction accuracy, because the model considers only the linear relationship of the historic database on the target road link, so the influence of upstream/downstream road link on the current road link has been neglected. When comparing the two prediction models with the real-time-data model, the real-time-data model has proven to be a worse predictor, possibly because the two prediction models can effectively discard unexpected information. Furthermore, the average MAPEs of the ANN model and k-NN model are close. Although the performance of the k-NN model is slightly better than the one of the ANN model, the ANN model is still an alternative method for the multi-time-step prediction of short-term traffic conditions. Finally, compared with the SVM model, the average MAPE of the k-NN model is slightly bigger, because the SVM model has better prediction accuracy. However, the generalizability performance of the SVM model is worse than that of the k-NN model, as the SVM model needs to re-optimize the parameters for different data sets. However, it is easier for k-NN models to be adapted to the prediction of short-term traffic conditions of other road links. Generally speaking, the prediction accuracy of the k-NN model can meet the demand of short-term traffic-flow prediction, and the advantages of the k-NN model are shown in the nonlinear trend of short-term traffic-condition prediction.

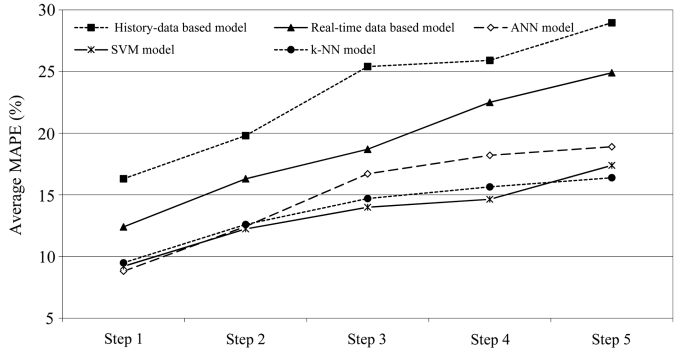


Fig. 15. Prediction accuracy comparison of the five models

Conclusions

This paper develops a k-NN model for the prediction of short-term traffic conditions. In the prediction model, a single-time-step prediction model, which synthetically considers the spatial and temporal parameters, is proposed first. Then, the multi-time-step prediction model is constructed based on an update to the single-time-step prediction model. To validate the performance of the proposed k-NN model, the GPS data of taxis in Foshan city are collected. Six road links on Foshan Avenue Road are used as the test bed, and the time interval for prediction is set as 5 min. The results show that the k-NN model has a better forecasting accuracy than the ANN model, real-time-data model, and history-data model. Although the forecasting accuracy of the k-NN model is slightly worse than that of the SVM model, the k-NN model is still very powerful considering both the forecasting accuracy and the generalizability performance. In addition, the prediction accuracy is observed to decrease with the increment of prediction steps. Overall, this study shows that the proposed k-NN model can obtain acceptable results with low MAPEs under the sophisticated road condition. Therefore, the k-NN-based multi-step model is a feasible way to forecast the short-term traffic condition. The proposed multi-time-step model can also provide more comprehensive traffic guidance for administrators and travelers.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (71571026, 51578112, and 51208079), the Trans-Century Training Program Foundation for Talents from the Ministry of Education of China (NCET-12-0752), Liaoning Excellent Talents in University (LJQ2012045), and the Fundamental Research Funds for the Central Universities (3013-852019 and 3132015062).

References

Akbari, M., van Overloop, P. J., and Afshar, A. (2011). "Clustered k nearest neighbor algorithm for daily inflow forecasting." *Water Resour. Manage.*, 25(5), 1341–1357.

Boto-Giralda, D., Diaz-Pernas, F. J., Gonzalez-Ortega, D., Diez-Higuera, J. F., Anton-Rodriguez, M., and Martinez-Zarzuela, M. (2010). "Wavelet-based denoising for traffic volume time series forecasting with self-organizing neural networks." *Comput.-Aided Civ. Infrastruct. Eng.*, 25(7), 530–545.

Chang, H., Lee, Y., Yoon, B., and Baek, S. (2012). "Dynamic near-term traffic flow prediction: System-oriented approach based on past experiences." *IET Intell. Transp. Syst.*, 6(3), 292–305.

Coric, V., Djuric, N., and Vucetic, S. (2012). "Traffic state estimation from aggregated measurements with signal reconstruction techniques." *Transp. Res. Rec.*, 2315, 121–130.

Jiang, X., Adeli, H. (2005). "Dynamic wavelet neural network model for traffic flow forecasting." *J. Transp. Eng.*, 10.1061/(ASCE)0733-947X (2005)131:10(771), 771–779.
Kamarianakis, Y., and Prastacos, P. (2003). "Forecasting traffic flow conditions in an urban network: Comparison of multivariate and univariate approaches." *Transp. Res. Rec.*, 1857, 74–84.
Kirby, H. R., Watson, S. M., and Dougherty, M. S. (1997). "Should we use neural networks or statistical models for short-term motorway traffic forecasting?" *Int. J. Forecasting*, 13(1), 43–50.
Lee, K. L., and Billings, S. A. (2003). "A new direct approach of computing multi-step ahead predictions for non-linear models." *J. Control*, 76(8), 810–822.
Li, R., and Rose, G. (2011). "Incorporating uncertainty into short-term travel time predictions." *Transp. Res. Part C*, 19(6), 1006–1018.
Lin, L., Wang, Q., and Sadek, A. K. (2013). "Short-term forecasting of traffic volume evaluating models based on multiple data sets and data diagnosis measures." *Transp. Res. Rec.*, 2392, 40–47.
Meng, M., Shao, C. F., and Wong, Y. D. (2015). "A two-stage short-term traffic flow prediction method based on AVL and AKNN techniques." *J. Central South Univ.*, 22(2), 779–786.
Min, W. L., and Wynter, L. (2011). "Real-time road traffic prediction with spatio-temporal correlations." *Transp. Res. Part C*, 19(4), 606–616.
Okutani, I., and Stephanedes, Y. J. (1984). "Dynamic prediction of traffic volume through Kalman filtering theory." *Transp. Res. Part B*, 18(1), 1–11.
Parlos, A. G., Rais, O. T., and Atiya, A. F. (2000). "Multi-step-ahead prediction using dynamic recurrent neural networks." *Neural Networks*, 13(7), 765–786.
Smith, B. L., and Demetsky, M. J. (1996). "Multiple-interval freeway traffic flow forecasting." *Transp. Res. Rec.*, 1554, 136–141.
Smith, B. L., and Demetsky, M. J. (1997). "Traffic flow forecasting: Comparison of modeling approaches." *J. Transp. Eng.*, 10.1061/(ASCE) 0733-947X(1997)123:4(261), 261–266.
Smith, B. L., Williams, B. M., and Oswald, R. K. (2002). "Comparison of parametric and nonparametric models for traffic flow forecasting." *Transp. Res. Part C*, 10(4), 303–321.
Tan, M. C., Wong, S. C., Xu, J. M., Guan, Z. R., and Zhang, P. (2009). "An aggregation approach to short-term traffic flow prediction." *Intell. Transp. Syst.*, 10(1), 60–69.
Turochy, R. E. (2006). "Enhancing short-term traffic forecasting with traffic condition information." *J. Transp. Eng.*, 10.1061/(ASCE)0733-947X (2006)132:6(469), 469–474.
Vlahogianni, E. I., Karlaftis, M. G., and Golias, J. C. (2007). "Spatio-temporal short-term urban traffic volume forecasting using genetically optimized modular networks." *Comput.-Aided Civ. Infrastruct. Eng.*, 22(5), 317–325.
Wei, Y., and Chen, M. C. (2012). "Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks." *Transp. Res. Part C*, 21(1), 148–162.
William, H. K., Tang, Y. F., and Tam, M. L. (2006). "Comparison of two non-parametric models for daily traffic forecasting in Hong Kong." *J. Forecasting*, 25(3), 173–192.
Xu, D. W., Dong, H. H., and Li, H. J. (2015). "The estimation of road traffic based on compressive sensing." *Transp. B-Transp. Dyn.*, 3(2), 131–152.
Yao, B. Z., Hu, P., Zhang, M. H., and Jin, M. Q. (2014a). "A support vector machine with the Tabu search algorithm for freeway incident detection." *Int. J. Appl. Math. Comput. Sci.*, 24(2), 397–404.
Yao, B. Z., Wang, Z., Zhang, M. H., Hu, P., and Yan, X. X. (2015). "Hybrid model for prediction of real-time traffic flow." *Proc. Inst. Civ. Eng. Transp.*, in press.
Yao, B. Z., Yao, J. B., Zhang, M. H., and Yu, L. (2014b). "Improved support vector machine regression in multi-step-ahead prediction for rock displacement surrounding a tunnel." *Scientia Iranica*, 21(4), 1309–1316.
Yoon, B., and Chang, H. (2014). "Potentialities of data-driven nonparametric regression in urban signalized traffic flow forecasting." *J. Transp. Eng.*, 10.1061/(ASCE)TE.1943-5436.0000662, 04014027.
Yu, B., Yang, Z. Z., Chen, K., and Yu, B. (2010). "Hybrid model for prediction of bus arrival times at next station." *J. Adv. Transp.*, 44(3), 193–204.
Zheng, W. Z., Lee, D. H., and Shi, Q. X. (2006). "Short-term freeway traffic flow prediction: Bayesian combined neural network approach." *J. Transp. Eng.*, 10.1061/(ASCE)0733-947X(2006)132:2(114), 114–121.
Zhong, H. L., Kuang, C. J., and Huang, X. Y. (2012). "Traffic flow prediction algorithm based on historical frequent pattern." *Comput. Eng. Des.*, 33(4), 1547–1552.
Zuo, W., Zhang, D., and Wang, K. (2008). "On kernel difference-weighted *k*-nearest neighbor classification." *Pattern Anal. Appl.*, 11(3–4), 247–257.