ZHAO LIU, Ph.D. Candidate[1]
E-mail: liuzhao_xy@sina.com
XIAO QIN, Ph.D.[2]
E-mail: qinx@uwm.edu
WEI HUANG, Ph.D.[1]
E-mail: seuhwei@126.com
XUANBING ZHU[3]
E-mail: zhuxuanbing@126.com
YUN WEI, Ph.D.[4]
E-mail: luckyboy0309@163.com
JINDE CAO, Ph.D.[5]
E-mail: jdcao@seu.edu.cn
JIANHUA GUO, Ph.D.[1]
(Corresponding Author)
E-mail: seugjh@163.com
[1] Intelligent Transportation System Research Center
Southeast University, Southeast University Road #2
Nanjing, 211189, P.R. China
[2] Civil and Environmental Engineering
University of Wisconsin-Milwaukee
Milwaukee, WI 53201-0784, USA
[3] Nanjing Foreign Language School
Beijing East Road #30, Nanjing, 210018, P.R. China
4 Beijing Urban Construction Design and Development
Group Co., Ltd, Fuchengmen North Street #5, Xicheng
District, Beijing, 100037, P.R. China
[5] School of Mathematics and Research Center for Complex
Systems and Network Sciences, Southeast University
Southeast University Road #2, Nanjing, 211189, P.R. China

Science in Traffic and Transport
Original Scientific Paper
Submitted: 8 Nov. 2017
Accepted: 25 Jan. 2019

EFFECT OF TIME INTERVALS ON K-NEAREST NEIGHBORS MODEL FOR SHORT-TERM TRAFFIC FLOW PREDICTION

ABSTRACT

The accuracy and reliability in predicting short-term traffic flow is important. The K-nearest neighbors (K-NN) approach has been widely used as a nonparametric model for traffic flow prediction. However, the reliability of the K-NN model results is unknown and the uncertainty of traffic flow point prediction needs to be quantified. To this end, we extended the K-NN approach by constructing the prediction interval associated with the point prediction. Recognizing the stochastic nature of traffic, time interval used to measure traffic flow rate is remarkably influential. In this paper, extensive tests have also been conducted after aggregating real traffic flow data into time intervals, ranging from 3 minutes to 30 minutes. The results show that the performance of traffic flow prediction can be improved when the time interval increases. More importantly, when the time interval is shorter than 10 minutes, K-NN can generate higher accuracy of the point prediction than the selected benchmark model. This finding suggests the K-NN model may be more appropriate for traffic flow point and interval prediction at a shorter time interval.

KEY WORDS

short-term traffic flow forecasting; point prediction; prediction interval; K-nearest neighbors; seasonal autoregressive integrated moving average (SARIMA); generalized autoregressive conditional heteroscedasticity (GARCH);

1. INTRODUCTION

The performance of intelligent transportation systems (ITS) largely depends on the accuracy and reliability of the real-time traffic information. Timely and prevalent traffic flow data enable the implementation of advanced traffic management strategies such as dynamic route guidance and adaptive signal control. In particular, short-term traffic flow prediction is fundamental for improving urban transportation systems operations, management and safety. Hence, many forecasting models have been developed and applied for predicting short-term traffic flow. In real world, the complexity of urban traffic due to large disparities among highway capacities and speeds, weather conditions, driver behavior and transportation policies constantly challenges the development of tangible and technically sound traffic flow forecasting models.

The traditional short-term traffic flow forecasting aims to predict an average value according to historical traffic patterns and trends. As there is always a degree of uncertainty associated with future point estimate, quantifying the reliability of the point prediction assures the expected range of future traffic fluctuations [1-2]. Although the accuracy of traffic prediction may remain the same, traffic managers and travelers can benefit from a reliability measurement by making informed decisions with calculated risk. Traffic flow interval prediction has been considered a key performance measure of transportation systems and hence methodologies have been proposed to quantify the reliability of the point prediction [3-7].

In short-term traffic flow forecasting, time interval serves as both the aggregation interval of traffic volume data and the forecasting horizon. Traffic flow rate calculation can be affected by the selection of time interval. Smith and Ulmer [8] investigated the impact of time interval on traffic flow series and found that traffic flow rate becomes stable with the increase of time interval. Guo et al. [1] investigated the effect of time intervals of traffic flow series on performance of traffic flow forecasting and found that the prediction methods needed further development for very short (e.g. less than 5 and 5-min) intervals. Overall, time interval is an important factor of determining the performance of short-term traffic forecasting.

Many methods have been proposed to predict short-term traffic condition. These methods can be generally classified into parametric or nonparametric. Parametric methods require a set of fixed parameters as part of their mathematical or statistical formulation. By contrast, nonparametric methods are generally driven by data and less constrained by the underlying model assumptions. Among parametric methods, statistical volatility models, e.g. the generalized autoregressive conditional heteroscedasticity (GARCH) model can quantify the uncertainty for short-term traffic forecasting through using a prediction interval [1, 4, 6, 9-11]. However, most of the nonparametric methods do not directly provide the uncertainty quantification for short-term traffic forecasting. As a nonparametric method, the K-nearest neighbors has been widely used to predict the point estimate for short-term traffic condition, but without interval prediction [12]. Targeting this gap, the objective of this study is to facilitate or enhance the K-NN method through quantifying the uncertainty of the point prediction, e.g. generating prediction interval for the K-NN method. To this end, comparative studies between statistical methods and the K-NN method will be conducted and the effect of time interval on forecasting performance will also be investigated for both point and interval prediction.

The remainder of the paper is organized as follows: after the introduction section, literature review is presented, followed by a methodology section. Next, a case study is presented. Finally, conclusions are summarized and future work is discussed.

2. LITERATURE REVIEW

The point prediction of short-term traffic conditions has been intensively investigated in the past decades. Since the late 1970s, parametric methods assumed that traffic condition data are linear stochastic in nature so that a linear structure can be applied to predict future traffic conditions. Typical parametric methods include historical average approach [13], autoregressive integrated moving average (ARIMA) models [14-16], Kalman filter method [7, 17-18] and spectral representation [19].

Unlike parametric methods, nonparametric methods are not restricted by the linearity assumption and, hence, nonlinear structure can be used to predict the future traffic condition. In general, nonparametric methods are driven by data. Typical nonparametric methods include neural network models [20-23], K-nearest neighbors [24-28], symbolic regression method [29] and support vector machine [30-32]. Refer to [33] for more prediction methods and additional details.

The uncertainty of traffic prediction can be described and measured through a prediction interval around the predicted average of the future traffic condition [34]. Guo et al. [37] validated the heteroscedastic nature of traffic conditional series. The statistical volatility method can be used to construct the prediction interval based on the conditional variance of traffic flow series [1, 4, 6, 9-11, 38]. There are a few methods for constructing prediction intervals for neural networks, such as the bootstrap, Bayesian and mean-variance prediction methods. Bootstrap is a resampling method that can yield prediction intervals with a high coverage probability [34-35]. The Bayesian method can be combined with neural networks for generating prediction intervals [5, 36]. The mean-variance prediction interval construction is based on the assumption that prediction error variance can be estimated through neural networks [2]. Compared with extensive research on point forecasting, uncertainty quantification is still at its infancy.

When forecasting short-term traffic condition, time interval plays an important role of determining the performance of short-term traffic flow prediction. The Highway Capacity Manual [39] recommended that a 15-min time interval can be used for most of the procedures and 5-min or less time intervals should be avoided. The majority of short-term traffic flow forecasting literature is focused on the development of prediction methods for time intervals ranging from 5-min to

15-min [7, 15, 19-20, 23, 25-27]. Also, some studies applied shorter time intervals (less than 5-min) [22, 24]. However, there is a limited number of studies on the selection of appropriate time intervals and their influence on the forecasting methods.

With the increasing demand for timely and proactive traffic control, traffic flow data with very short time intervals became more important. A comprehensive investigation is necessary for pairing forecasting methods with time intervals to individual applications such that robustness, accuracy and practical utility can be maximized [1]. Reliable short-term traffic condition forecasting requires the provision of prediction interval to quantify the uncertainty. Most of the literature on short-term traffic flow forecasting focus on improving the performance of the point prediction and the quantification of uncertainty associated with the point prediction is still to be investigated. The K-NN method is a proven method for the point prediction but has not been applied to generate the prediction interval. This paper extended the K-NN method to quantify the uncertainty of the point prediction and evaluated the impact of time intervals accordingly.

3. METHODOLOGY

3.1 State vector definition for the K-NN method

The fundamental assumption of K-NN is that future states are similar to their neighbors of the past [25]. For a specific time interval, a classical state vector defines the collected traffic flow data at times t, t-1, ... , t-d, where d is an appropriate number of lags. For instance, Smith et al. [25] defined a hybrid state vector that included three lagged observations and two historical averages for the 15-min traffic flow data. The hybrid state vector was proven to be reasonable in terms of sufficiency and simplicity. The same definition of a hybrid state vector was adopted as shown in *Table 1*. In *Table 1*, $V_c(t)$ is the traffic flow at current time t and $V_{hist,c}(t)$ is the historical average of traffic flow at the time of day and day of week associated with time interval t. Considering the stability of traffic flow series at shorter and longer time intervals, the lagged observations at shorter time intervals were added and the lagged observations at the longer time interval were reduced. Specifically, three state vectors for different time intervals ranging from 3-min to 30-min were defined.

In order to select the best neighbors from the historical data, it is necessary to measure the degree of proximity between two state vectors. In this study, Euclidean distance was applied due to its simplicity and effectiveness.

$$d_i = \| X_t - X_i \|, \quad i = 1,2,3,\ldots,N \tag{1}$$

where X_t denotes the current state vector and X_i denotes the i-th state vector in the historical database. The K historical state vector samples with the least distance to the current state vector are selected to generate the forecasts.

3.2 The point prediction using K-NN

According to the distance between the current and historical state vectors, the ranked K-nearest neighbors can be directly used to calculate point traffic flow predictions. For example, straight average of the selected neighbors was used for predicting the traffic flow in Smith et al. [25]. In this paper, four forecast functions are formulated in *Equations 2-5*.

Straight average:

$$\hat{V}(t+1) = \frac{1}{k}\sum_{i=1}^{k} V_i(t+1) \tag{2}$$

Adjusted by $V(t)$:

$$\hat{V}(t+1) = \frac{1}{k}\sum_{i=1}^{k} V_i(t+1)\frac{V_c(t)}{V_i(t)} \tag{3}$$

Adjusted by $V_{hist}(t+1)$:

$$\hat{V}(t+1) = \frac{1}{k}\sum_{i=1}^{k} V_i(t+1)\frac{V_{hist,c}(t)}{V_{hist,i}(t)} \tag{4}$$

Adjusted by both $V(t)$ and $V_{hist}(t+1)$:

$$\hat{V}(t+1) = \frac{1}{k}\sum_{i=1}^{k} V_i(t+1)\left[\frac{V_c(t)}{V_i(t)} + \frac{V_{hist,c}(t)}{V_{hist,i}(t)}\right]\cdot\frac{1}{2} \tag{5}$$

where $\hat{V}(t+1)$ is the future traffic flow at the next time interval, $V_i(t)$ is the traffic flow value in the i-th historical state vector at time t, k is the number of selected nearest neighbors, $V_c(t)$ is the traffic flow at the current time, $V_{hist,c}(t)$ is the historical traffic flow observed at the current time and $V_{hist,i}(t)$ is the traffic flow in the i-th historical state vector.

3.3 Prediction interval construction

The conventional application of K-NN is to generate point future traffic flow which does not carry information regarding the level of confidence. According to the

Table 1 – Definition of state vector for different time intervals

Time intervals	Defined state vector
T ≤ 10 min	$[V_c(t), V_c(t\text{-}1), V_c(t\text{-}2), V_c(t\text{-}3), V_c(t\text{-}4), V_{hist,c}(t\text{-}2), V_{hist,c}(t\text{-}1), V_{hist,c}(t), V_{hist,c}(t+1)]$
10 min < T ≤ 20 min	$[V_c(t), V_c(t\text{-}1), V_c(t\text{-}2), V_c(t\text{-}3), V_{hist,c}(t\text{-}1), V_{hist,c}(t), V_{hist,c}(t+1)]$
T > 20 min	$X(t)=[V_c(t), V_c(t\text{-}1), V_c(t\text{-}2), V_{hist,c}(t), V_{hist,c}(t+1)]$

point prediction based on conventional K-NN, when a series of traffic flow data (v_1, v_2, ..., v_n) are observed at times (t_1, t_2, ..., t_n), future traffic flow can be estimated by processing the selected neighbors at each time. Clearly, the prediction value changes over time and the number and the composition of the neighbors will also change accordingly. Therefore, for each point value predicted, the time-dependent neighbors from the historical traffic data at the same time can be chosen to construct the prediction interval. *Figure 1* illustrates the relationship between the point prediction and the prediction interval.

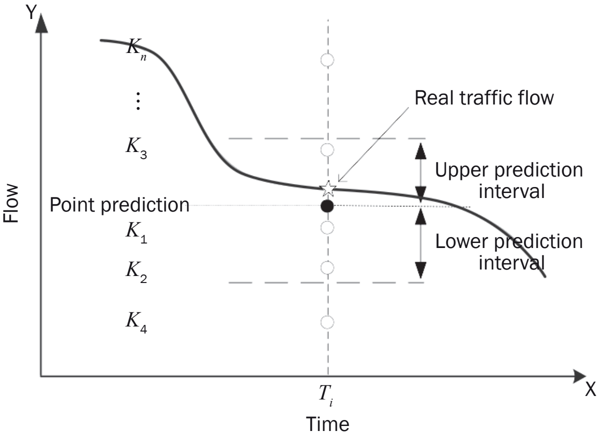


Figure 1 – Relationship between target v_t and prediction interval

Figure 1 shows that at a specific time T_i, n nearest neighbors Ks (denoted by the gray dots) can generate a predicted traffic flow (i.e., point prediction), denoted by the black dot. Based on the selected nearest neighbors of Ks, a prediction interval consisting of upper and lower prediction limits can be provided to predict the range of traffic flow oscillation during the next time interval. At a specific time T_i the prediction interval for the point prediction can be computed using the variance of the selected nearest neighbors as *Equation 6*.

$$\hat{\sigma}_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \tag{6}$$

where n_i is the number of the selected neighbors at time T_i, y_{ij} is the j-th selected neighbors at time T_i and $\bar{y}_i$ is the predicted traffic flow at the time T_i. The (1-α) prediction interval or confidence interval at T_i can be formulated as *Equation 7*.

$$\frac{\bar{y}_i \pm t(n_i - 1, \alpha)\hat{\sigma}_i}{\sqrt{n_i}} \tag{7}$$

where $t(h,\alpha)$ denotes the α-th quantile of the Student's t-distribution on h degrees of freedom. The prediction interval should contain the future traffic flow at the given confidence level and the width of the prediction interval indicates the degree of the uncertainty.

3.4 Prediction performance measures

The performance of the point prediction can be evaluated by the mean absolute percentage error (MAPE), mean absolute error (MAE) and root mean square error (RMSE), and smaller values of these three measures indicate better model performance.

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{X_i - \hat{X}_i}{X_i} \right| \tag{8}$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} | X_i - \hat{X}_i | \tag{9}$$

$$RMSE = \frac{1}{n} \sqrt{n \sum_{i=1}^{n} (X_i - \hat{X}_i)^2} \tag{10}$$

where X_i denotes the real-world observation, X_i stands for the forecasts and n is the total number of observations $\hat{X}_i$.

For interval prediction performance evaluation, Guo et al. [1] proposed two guiding principles: (a) real observations should fall within the prediction intervals with respect to the selected confidence level and (b) narrower predicted intervals, if valid, are more informative than the wider ones. Therefore, prediction interval kickoff percentage and prediction interval width-to-flow ratio are used to evaluate interval prediction. Their formulations are expressed in *Equations 11* and *12*, respectively.

$$kickoff = \frac{1}{n} \sum_{i=1}^{n} count_i, \quad count_i = \begin{cases} 1, & \hat{X}_i^{low} > X_i, \ or, \ X_i > \hat{X}_i^{up} \\ 0, & \hat{X}_i^{low} < X_i < \hat{X}_i^{up} \end{cases} \tag{11}$$

$$Width = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{X}_i^{up} - \hat{X}_i^{low}}{X_i} \tag{12}$$

where X_i denotes the actual observations, X_i^{low} stands for the forecasts of the lower bound, X_i^{up} stands for the forecasts of the upper bound and n is the total number of observations. For the 95% confidence interval, the kickoff percentage should be close to 5%. Meanwhile, the width-to-flow ratio should be as small as possible.

4. EMPIRICAL STUDY

In this section, an empirical study was conducted using the traffic data from three highway locations around London, UK. The number of nearest neighbors for the K-NN method was identified, based on which, both the point prediction and the prediction interval were computed and compared with the benchmark method of seasonal autoregressive integrated moving average (SARIMA) plus GARCH, i.e., SARIMA + GARCH.

4.1 Data description

The traffic flows used in this study were from Station 4762A in the southwest of the M25 motorway, Stations 2737A and 2808B in the northwest of the M1 motorway, London, UK. The traffic data were collected by Highways Agency through the Motorway Incident Detection and Automatic Signaling (MIDAS) system which

Table 2 – Descriptive information for the collected stations

Region	Highway	Stations	Number of lanes	Start	End	Simple size
UK	M25	4762a	4	1/1/2002	12/31/2002	35,040
UK	M1	2737a	3	2/13/2002	12/31/2002	30,912
UK	M1	2808b	3	2/13/2002	12/31/2002	30,912

archives traffic condition data on a 1-min interval, 24 h per day and 7 days per week. The descriptive information for the stations is listed in *Table 2*. The original 1-min traffic data for each station were then aggregated into traffic flow series with different time intervals (i.e., 3, 5, 7, 10, 12, 15, 18, 20, 24, 28 and 30 min).

4.2 Study design

This study was designed to evaluate the proposed prediction intervals based on K-NN and to investigate the effect of different time intervals on the performance of prediction models. The performance of the point prediction of selected forecasting models was measured for each of the 11 time intervals. The SARIMA + GARCH model was used as a benchmark.

For K-NN, data collected for the first nine months (from January/February to September) were used as historical data and data for the last three months (from October to December) were used as test data. For SARIMA + GARCH, weekly seasonal period was used. Different from K-NN, SARIMA + GARCH does not require a historical database. Therefore, the data from the last week in September was used to activate the SARIMA + GARCH model.

4.3 Identification of the number of neighbors K

It is important to use the appropriate number of nearest neighbors to improve both point prediction and prediction interval. A wide range of nearest neighbors was considered, ranging from one candidate to fifty candidates. For brevity, the straight average forecast method was used to present the number of nearest neighbors for Station 2737a. In *Figure 2*, it is clear that the number of neighbors affects the accuracy of the point prediction. A large decrease in the forecast error can be observed before the number of neighbors reaches 10; afterwards, the forecast error gradually declines and converges. In addition, the forecast error decreases clearly with the increase of time interval, indicating that the stability and predictability of traffic flow are positively correlated.

The model performance pertaining to the number of neighbors and interval length for the prediction interval at 95% confidence level was calculated. In *Figure 3a*, the kickoff percentage rapidly ascends from around 5% to 75% with the increase of the number of nearest neighbors, and the ascending rate is higher from 0 to 10 nearest neighbors and decreases afterwards. In *Figure 3b*, the width-to-flow ratio rapidly descends when the number of neighbors is approaching 4 and then becomes stable. Both observations show the decreased prediction performance with the increase of the number of nearest neighbors.

Table 3 lists the optimal number of nearest neighbors for point and interval prediction. For all the four forecasting models, the number of nearest neighbors decreases gradually with the increase of time interval. The number of nearest neighbors is minimum when the time interval is 30 minutes. For each of the four models, the minimum numbers of nearest neighbors at 30-min time interval are 10, 30, 6 and 10, respectively. In contrast, only two nearest neighbors are needed to generate workable prediction intervals for all four forecast functions across all tested time intervals.

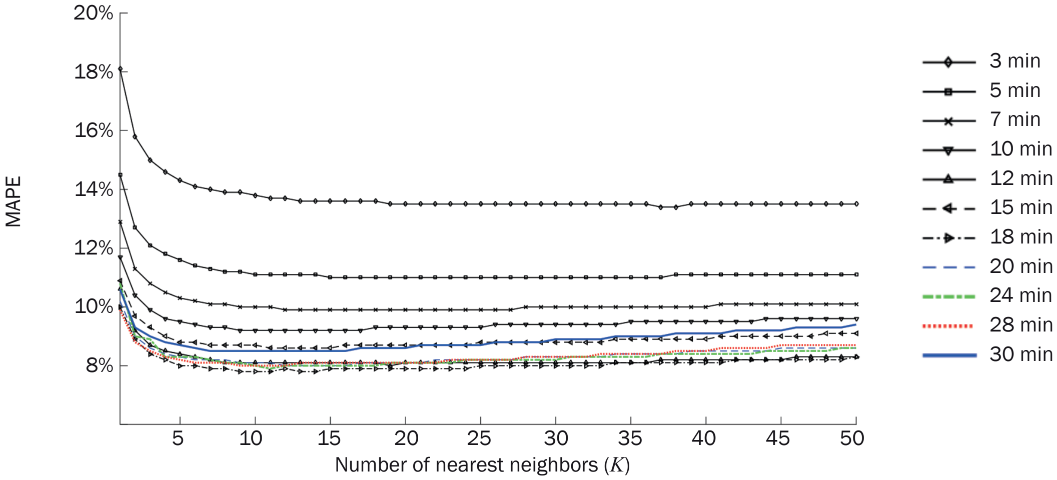
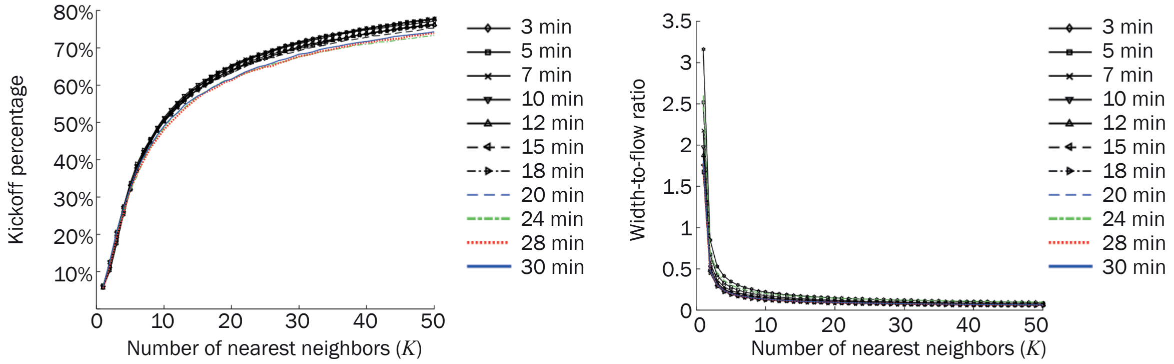


Figure 2 – The point prediction MAPE at 2737a



a)

b)

Figure 3 – Interval prediction performance at 95% confidence interval at 2737a

Table 3 – Determination of the optimal number of nearest neighbors

Time interval [min]	Straight average		$V(t)$		$V_{hist}(t+1)$		$V(t)$ and $V_{hist}(t+1)$	
	Point	Interval	Point	Interval	Point	Interval	Point	Interval
3	40	2	35	2	23	2	35	2
5	30	2	30	2	20	2	27	2
7	23	2	30	2	13	2	24	2
10	15	2	22	2	12	2	16	2
12	21	2	25	2	12	2	22	2
15	13	2	25	2	10	2	14	2
18	11	2	25	2	9	2	15	2
20	15	2	30	2	10	2	14	2
24	12	2	17	2	8	2	12	2
28	10	2	17	2	6	2	10	2
30	10	2	30	2	6	2	10	2

4.4 The point prediction comparison

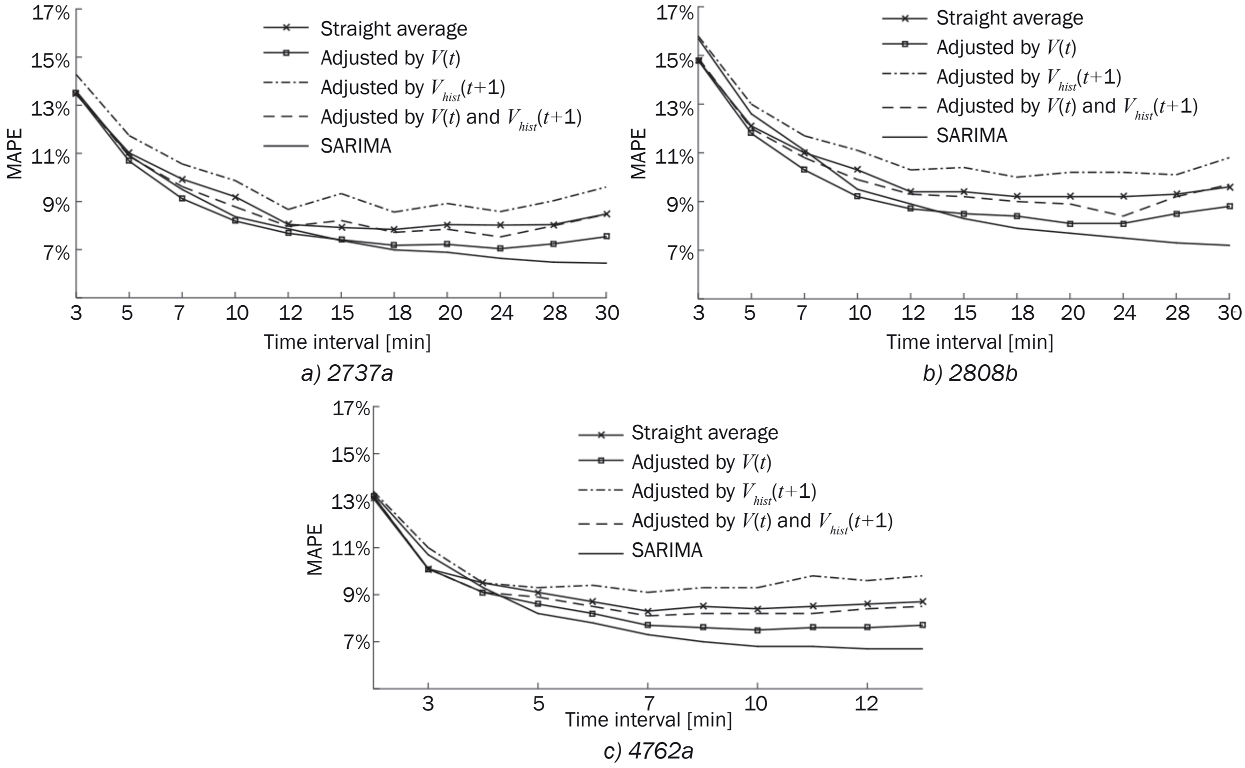
After determining the number of nearest neighbors for K-NN, the comparison of the point prediction between K-NN and benchmark models is presented in *Figures 4-6*, respectively. In *Figure 4*, the MAPEs of SARIMA gradually decrease with the increase of time interval and the four forecast models of K-NN gradually decrease before the time interval reaches 24 minutes. Specifically, for Station 2737a (*Figure 4a*) and Station 2808b (*Figure 4b*), when the time interval is less than 15 minutes, the forecast model Adjusted by $V(t)$ of K-NN is better than SARIMA and the other three K-NN forecast models. When the time interval is between 15 to 30 minutes, SARIMA gradually outperforms the rest. Similarly, for Station 4762a (*Figure 4c*), K-NN shows better performance than SARIMA before the time interval reaches 10 minutes, while SARIMA outperforms K-NN when the time interval is between 10 and 30 minutes. It is evident that the K-NN model is more appropriate for traffic flow forecasting for shorter time intervals and the SARIMA model is more appropriate for longer time intervals.

For MAE and RMSE, the results are shown in *Figures 5* and *6*, respectively, with similar patterns as demonstrated in *Figure 4*. Specifically, for shorter time intervals, K-NN can obtain better performance of the point prediction than the SARIMA model. Conversely, the SARIMA model outperformed K-NN for longer time intervals.

4.5 Prediction interval comparison

In addition to the point prediction, the K-NN models can obtain prediction interval of traffic flow. Given a 95% confident interval, the measures of kickoff percentage and width-to-flow ratio for prediction interval coverage are presented in *Figures 7* and *8*, respectively.

Recall that kickoff percentage of prediction intervals should be close to 5%, for 95% confidence level. *Figure 7* shows that at 95% confidence level, the kickoff percentages from four K-NN forecast models are all around 5% for all the time intervals and across all the three stations. Comparatively, the GARCH model has a 2% kickoff percentage at Stations 2737a and 2808b, and a 2.5% kickoff at Station 4762a. Therefore, four K-NN models are better than GARCH in terms of kickoff percentage. In addition, *Figure 8* shows that the width-to-flow ratios for GARCH model declines to 0.8 which is much smaller than K-NN with the values descending from 3.5 to 2. Therefore, the performance of prediction interval generation of the K-NN model is mixed compared with that of the conventional GARCH model.

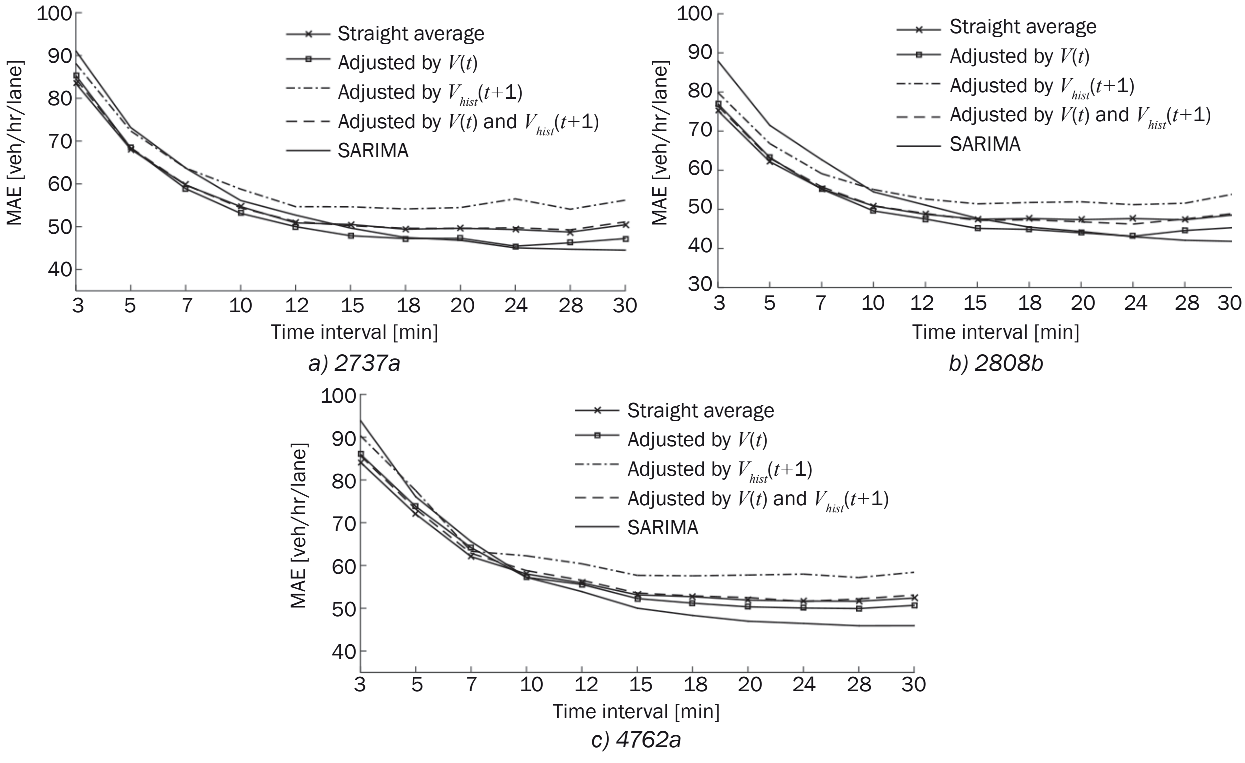


a) 2737a

b) 2808b

c) 4762a

Figure 4 – MAPE comparison of the point prediction

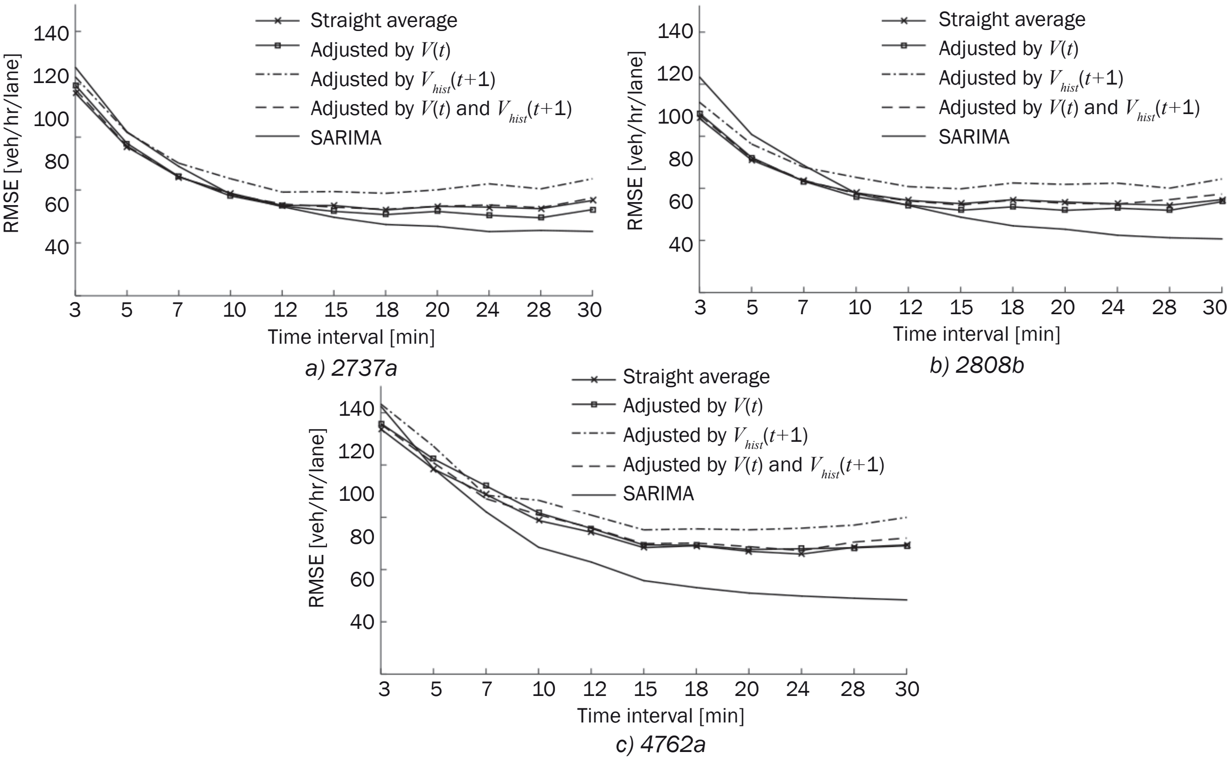


a) 2737a

b) 2808b

c) 4762a

Figure 5 – MAE comparison of the point prediction

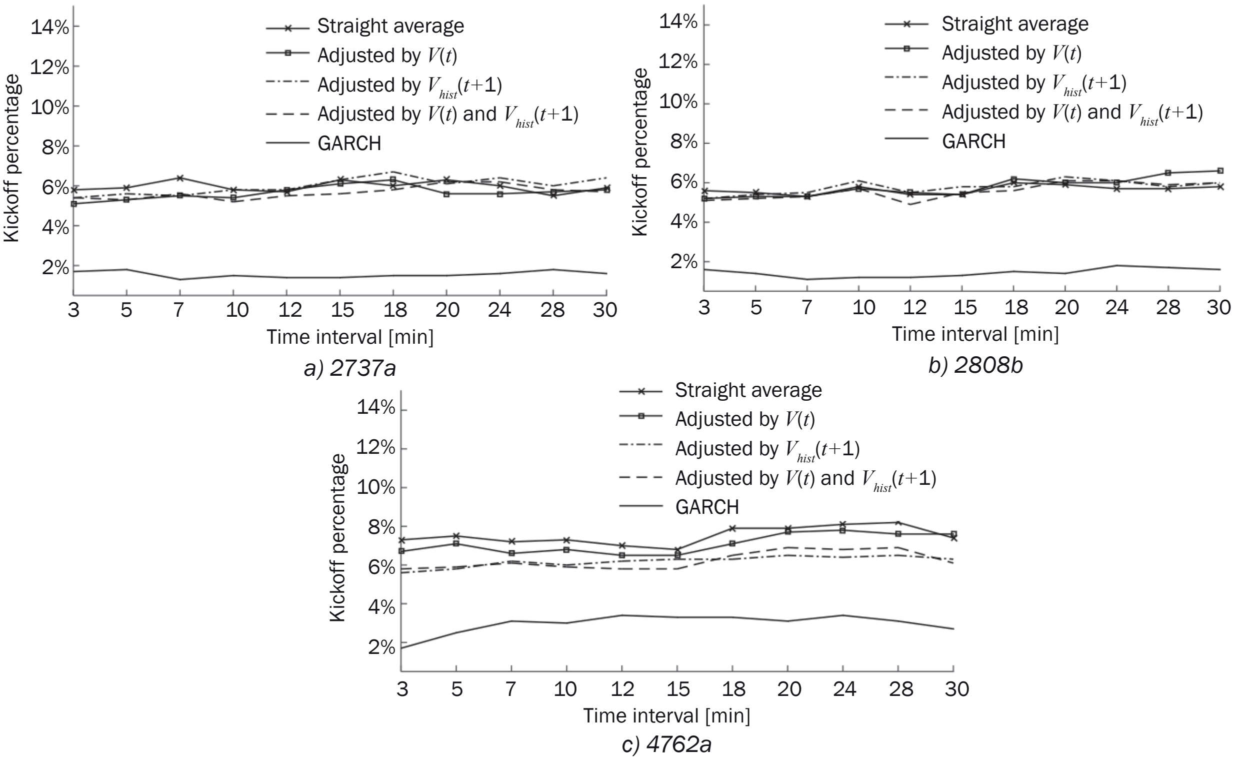


a) 2737a

b) 2808b

c) 4762a

Figure 6 – RMSE comparison of the point prediction

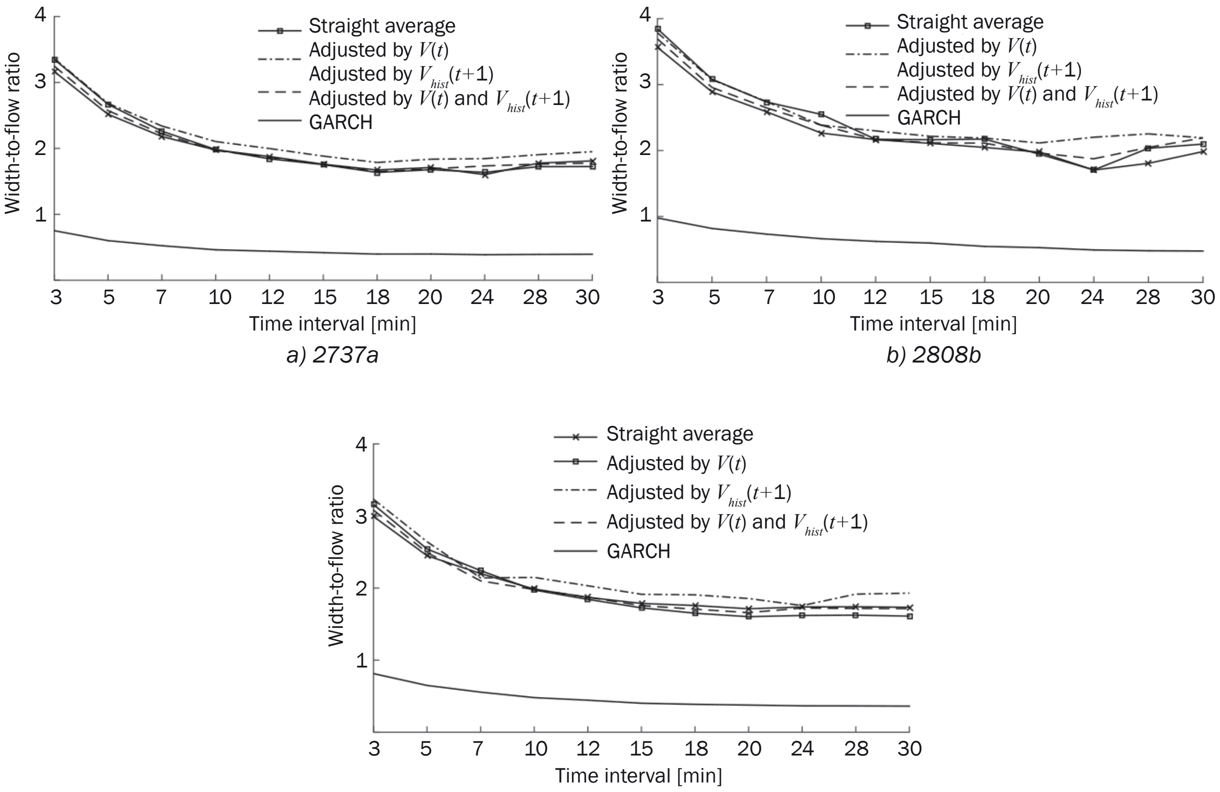


a) 2737a

b) 2808b

c) 4762a

Figure 7 – Kickoff percentage comparison of prediction interval under 95% confidence interval



a) 2737a

b) 2808b

c) 4762a

Figure 8 – Width-to-flow ratio comparison of prediction interval under 95% confidence interval

5. CONCLUSIONS

Short-term traffic flow forecasting models are crucial for supporting proactive traffic control and management. Desirable short-term traffic flow forecasting should include both the point prediction of traffic flow and its prediction interval for measuring the prediction uncertainty. Due to the stochastic nature of traffic flow, the prediction interval is increasingly important to traffic managers and travelers who prefer to make decisions under known uncertainties.

This study extends the K-NN model to construct the prediction interval at the 95% confidence level. As a nonparametric regression model, the K-NN model is compared to the benchmark SARIMA + GARCH model. In an empirical study, real-world traffic flow data were aggregated into various time intervals ranging from 3 to 30 minutes. Extensive tests have been conducted for both K-NN and SARIMA + GARCH models. The results provide consistent findings pertinent to the effect of time interval on the modeling performance of the point prediction and the prediction interval between the two methods. The investigation of the point prediction shows that K-NN can generate higher accuracy when a time interval is shorter than 10 minutes; however, the performance of prediction interval of K-NN is mixed compared with the conventional GARCH model. As opposed to fitting and optimizing parametric models (e.g. SARIMA), the main advantage of the K-NN model is that the predictions are generated from the historical pattern of traffic flow, without being restricted by the linearity assumption of traffic data. However, compared to the GARCH model, the K-NN model in this study did not consider the time-dependent correlation of prediction intervals.

Future studies are recommended to further develop the prediction interval of short-term traffic flow forecasting. In addition, future studies can also be conducted in accommodating the generated prediction interval in transportation management and control applications. For example, the prediction interval of traffic flow rate can be used in signal timing to improve the reliability of road network traffic signal control.

ACKNOWLEDGEMENT

The authors thank the United Kingdom Highways Agency for providing the traffic data used in this research. The authors hold all the responsibility with respect to the analyses and views presented in this work. This study is supported by the National Natural Science Foundation of China (Grant No. 61573106) and the Jiangsu Provincial Key Laboratory of Networked Collective Intelligence (Grant No. BM2017002).

刘钊，秦晓，黄卫，朱炫冰，魏运，曹进德，郭建华*

时间间隔对短时交通流K近邻预测模型的影响

摘要

短期交通流预测的准确性和可靠性是非常重要的。K近邻方法(*K-NN*)作为一种非参数交通流预测模型已得到了广泛的应用。然而，*K-NN*模型预测结果的可靠性没有得到足够的关注，交通流点预测的不确定性需要量化。为此，我们对*K-NN*方法进行了扩展，构造了与点预测相关联的预测区间。考虑交通流的随机性，用来汇集交通流的时间间隔具有十分重要的影响作用。在本文的实验分析中，交通流数据将按照不同的时间间隔进行汇集从而进行分析，时间间隔从*3*分钟到*30*分钟。结果表明，交通流预测的精度随着时间间隔的增加而增加。更重要的是，当时间间隔小于*10*分钟时，相比较基准模型，*K-NN*可以产生更高的点预测精度。这表明*K-NN*模型可能更适合于较短时间间隔下的交通流点预测和区间预测。

关键词

短时交通流预测；点预测；预测区间；K近邻方法；季节性差分自回归移动平均模型*(SARIMA)*；广义自回归条件异方差模型*(GARCH)*

REFERENCES

[1] Guo J, Williams B, Smith B. Data collection time intervals for stochastic short-term traffic flow forecasting. *Transportation Research Record: Journal of the Transportation Research Board.* 2008;(2024): 18-26.

[2] Mazloumi E, Rose G, Currie G, Moridpour, S. Prediction intervals to account for uncertainties in neural network predictions: Methodology and application in bus travel time prediction. *Engineering Applications of Artificial Intelligence*. 2011;24(3): 534-542.

[3] Pattanamekar P, Park D, Rilett LR, Lee J, Lee C. Dynamic and stochastic shortest path in transportation networks with two components of travel time uncertainty. *Transportation Research Part C: Emerging Technologies.* 2003;11(5): 331-354.

[4] Tsekeris T, Stathopoulos A. Real-time traffic volatility forecasting in urban arterial networks. *Transportation Research Record: Journal of the Transportation Research Board.* 2006;1964: 146-156.

[5] Khosravi A, Nahavandi S, Creighton D. A prediction interval-based approach to determine optimal structures of neural network metamodels. *Expert systems with applications.* 2010;37(3): 2377-2387.

[6] Chen C, Hu J, Meng Q, Zhang Y. Short-time traffic flow prediction with ARIMA-GARCH model. In: *2011 IEEE Intelligent Vehicles Symposium (IV).* IEEE; 2011. p. 607-612.

[7] Guo J, Huang W, Williams BM. Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification. *Transportation Research Part C: Emerging Technologies.* 2014;43: 50-64.

[8] Smith BL, Ulmer JM. Freeway traffic flow measurement: investigation into impact of measurement time interval. *Journal of Transportation Engineering.* 2003;129(3): 223-229.

[9] Sohn K, Kim D. Statistical model for forecasting link travel time variability. *Journal of Transportation Engineering.* 2009;135(7): 440-453.

[10] Yang M, Liu Y, You Z. The reliability of travel time forecasting. *IEEE Transactions on Intelligent Transportation Systems.* 2010;11(1): 162-171.

[11] Zhang Y, Sun R, Haghani A, Zeng X. Univariate volatility-based models for improving quality of travel time reliability forecasting. *Transportation Research Record: Journal of the Transportation Research Board.* 2013;2365: 73-81.

[12] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*. 1992;46(3): 175-185.

[13] Williams B, Durvasula P, Brown D. Urban freeway traffic flow prediction: application of seasonal autoregressive integrated moving average and exponential smoothing models. *Transportation Research Record: Journal of the Transportation Research Board.* 1998;1644: 132-141.

[14] Hamed MM., Al-Masaeid HR, Said ZMB. Short-term prediction of traffic volume in urban arterials. *Journal of Transportation Engineering.* 1995;121(3): 249-254.

[15] Williams BM, Hoel LA. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. *Journal of transportation engineering.* 2003;129(6): 664-672.

[16] Lippi M, Bertini M, Frasconi P. Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. *IEEE Transactions on Intelligent Transportation Systems.* 2013;14(2): 871-882.

[17] Okutani I, Stephanedes YJ. Dynamic prediction of traffic volume through Kalman filtering theory. *Transportation Research Part B: Methodological.* 1984;18(1): 1-11.

[18] Wang Y, Papageorgiou M. Real-time freeway traffic state estimation based on extended Kalman filter: a general approach. *Transportation Research Part B: Methodological.* 2005;39(2): 141-167.

[19] Zhang Y, Zhang Y, Haghani A. A hybrid short-term traffic flow forecasting method based on spectral analysis and statistical volatility model. *Transportation Research Part C: Emerging Technologies*. 2014;43: 65-78.

[20] Dougherty MS, Cobbett MR. Short-term inter-urban traffic forecasts using neural networks. *International Journal of Forecasting.* 1997;13(1): 21-31.

[21] Yun SY, Namkoong S, Rho JH., Shin SW, Choi JU. A performance evaluation of neural network models in traffic volume forecasting. *Mathematical and Computer Modelling.* 1998;27(9-11): 293-310.

[22] Vlahogianni EI, Karlaftis MG, Golias JC. Optimized and meta-optimized neural networks for short-term traffic flow prediction: A genetic approach. *Transportation Research Part C: Emerging Technologies.* 2005;13 (3): 211-234.

[23] Kumar K, Parida M, Katiyar VK. Short term traffic flow prediction in heterogeneous condition using artificial neural network. *Transport.* 2015;30(4): 397-405.

[24] Davis GA, Nihan NL. Nonparametric regression and short-term freeway traffic forecasting. *Journal of Transportation Engineering.* 1991;117(2): 178-188.

[25] Smith BL, Williams BM, Oswald RK. Comparison of parametric and nonparametric models for traffic flow

forecasting. *Transportation Research Part C: Emerging Technologies.* 2002;10(4): 303-321.

[26] Zheng Z, Su D. Short-term traffic volume forecasting: A k-nearest neighbor approach enhanced by constrained linearly sewing principle component algorithm. *Transportation Research Part C: Emerging Technologies.* 2014;43: 143-157.

[27] Habtemichael FG, Cetin M. Short-term traffic flow rate forecasting based on identifying similar traffic patterns. *Transportation Research Part C: Emerging Technologies.* 2016;66: 61-78.

[28] Cai P, Wang Y, Lu G, Chen P, Ding C, Sun J. A spatiotemporal correlative k-nearest neighbor model for short-term traffic multistep forecasting. *Transportation Research Part C: Emerging Technologies.* 2016;62: 21-34.

[29] Li L, Tomislav F, Zhang J, Ran B. Traffic Speed Prediction for Highway Operations Based on a Symbolic Regression Algorithm. *Promet - Traffic & Transportation.* 2017;29(4); 433-441.

[30] Zhang Y, Xie Y. Forecasting of short-term freeway volume with v-support vector machines. *Transportation Research Record: Journal of the Transportation Research Board.* 2008;2024: 92-99.

[31] Peng T, Tang Z. A small scale forecasting algorithm for network traffic based on relevant local least squares support vector machine regression model. *Applied Mathematics & Information Sciences.* 2015;9(2L): 653-659.

[32] Li L, He S, Zhang J, Ran B. Short-term highway traffic flow prediction based on a hybrid strategy considering temporal-spatial information. *Journal of Advanced Transportation.* 2016;50(8): 2029-2040.

[33] Vlahogianni EI, Golias JC, Karlaftis MG. Short-term traffic forecasting: overview of objectives and methods. *Transport Reviews*. 2004;24 (5): 533-557.

[34] Heskes T. Practical confidence and prediction intervals. In: Mozer TPM, Jordan M. (Eds.) *Neural Information Processing Systems.* Cambridge, MA: MIT Press. 1997;9: 176-182.

[35] Rivals I, Personnaz L. Construction of confidence intervals for neural networks based on least squares estimation. *Neural Networks*. 2000;13(4-5): 463-484.

[36] Van Hinsbergen CI, Van Lint JWC, Van Zuylen HJ. Bayesian committee of neural networks to predict travel times with confidence intervals. *Transportation Research Part C: Emerging Technologies.* 2009;17(5): 498-509.

[37] Guo J, Huang W, Williams BM. Integrated heteroscedasticity test for vehicular traffic condition series. *Journal of Transportation Engineering.* 2012;138(9): 1161-1170.

[38] Guo J, Williams B. Real-time short-term traffic speed level forecasting and uncertainty quantification using layered Kalman filters. *Transportation Research Record: Journal of the Transportation Research Board.* 2010;2175: 28-37.

[39] Transportation Research Board TRB. *Highway Capacity Manual.* Transportation Research Circular–Special Rep. 209, National Research Council, Washington, D.C, 1998. Available from: ftp://public-ftp.agl.faa.gov/OMP%20PFC%2006-19-C--00-ORD/EIS%20and%20ROD%20Administrative%20Record/Disk01/!1918-1999/1997/11_99_1257.pdf