View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

ERDEM DOĞAN, Ph.D.
E-mail: edogan@kku.edu.tr
Kırıkkale University, Engineering Faculty
Department of Civil Engineering
Kirikkale, Yahsihan, 71451, Turkey

Intelligent Transport Systems (ITS)
Original Scientific Paper
Submitted: 6 Jan. 2019
Accepted: 3 Oct. 2019

SHORT-TERM TRAFFIC FLOW PREDICTION USING ARTIFICIAL INTELLIGENCE WITH PERIODIC CLUSTERING AND ELECTED SET

ABSTRACT

Forecasting short-term traffic flow using historical data is a difficult goal to achieve due to the randomness of the event. Due to the lack of a solid approach to short-term traffic prediction, the researchers are still working on novel approaches. This study aims to develop an algorithm that dynamically updates the training set of models in order to make more accurate predictions. For this purpose, an algorithm called Periodic Clustering and Prediction (PCP) has been developed for use in short-term traffic forecasting. In this study, PCP was used to improve Artificial Neural Networks (ANN) predictive performance by improving the training set of ANN to predict short-term traffic flow using selected clusters. A large amount of traffic data collected from the US and UK motorways was used to determine the PCP ability to increase the ANN performance. The robustness of the proposed approach was determined by the performance measures used in the literature and the mean prediction errors of PCP were significantly below other approaches. In addition, the studies showed that the percentage errors of PCP predictions decreased in response to increasing traffic flow values. Considering the obtained positive results, this method can be used in real-time traffic control systems and in different areas needed.

KEY WORDS

traffic prediction; training set; short-term prediction; k-means; artificial neural networks;

1. INTRODUCTION

Road traffic has become a more difficult event to manage as a result of the growth of cities and the increasing demand for transportation. Therefore, the development of systems that can effectively manage this complex event has become an extremely important issue today. Attempts are made to direct the strategies applied in traffic management and control systems to the traffic effectively. However, reliable predictions of traffic variables are required for these systems to work effectively. These variables include traffic flow, travel time, speed, intensity, occupancy, etc. As a result of reliable short-term forecasts the traffic flows can be controlled dynamically, consistent strategies for emergencies can be developed and signal systems can be optimized.

In order to assist the traffic management, the researchers have developed many methods with the motivation of accurately predicting the traffic flow. As a result of these studies, it was observed that the performance of the approaches decreased when the time resolution of traffic flow data increased. For this reason, a fully accepted approach to short-term traffic flow prediction has not been developed yet and the development studies are still underway. Moreover, the time horizon of the methods varies from 0.1 minutes to 1 day [1] and most researchers are developing models using datasets in different time horizons. For this reason, it is difficult to make a reliable comparison between approaches. Therefore, it is important to develop models by using a sufficient set of data to increase the effectiveness of the models. From this point of view, attention was paid to ensure that the datasets used in this study were of sufficient variety and quantity. A dataset containing traffic flow rates, which is first used in [2], and then in [3] was obtained from researchers and used in this study. The datasets used in this study were of varied sizes (3 months to 12 months) and were obtained from freeways and motorways of the United Kingdom and the United States. These datasets are superior in diversity and size compared to the relevant studies in the literature.

Today, many new hybrid models are being developed. The efficacy and stable operating conditions of these models have not been fully proven. However, Artificial Neural Networks (ANNs) and k-means algorithms are stable algorithms that have proven effective in the literature. Therefore, it was decided to run PCP together with ANNs and k-means algorithms in order to clearly observe the effectiveness of PCP. In data-driven models, the suitability of samples within the training cluster to the problem improves the performance of the model. Thus, it is ensured that the approaches such as ANNs, which contain randomness, can make more accurate predictions more frequently. The main motivation of the developed method is to prepare the training sets to improve the predictive performance of models that can be trained with datasets with consecutive data samples. For this purpose, the Periodic Clustering and Prediction (PCP) has been developed with a novel approach of k-means clustering algorithm and ANNs. The general procedure of PCP is as follows. First, k-means clustering algorithm was used to divide the main training set into subsets. This processing step is performed to identify past traffic flow patterns that have previously resulted in similar traffic flow values. Subsequently, the elected set $\{e^*\}$ is determined from the subsets. The e^* is the training set that the model uses to estimate the value of the future traffic flow and contains the most appropriate samples for the estimation. Thus, the model does not only use a certain amount of historical data in an ordered dataset, but it is trained with an appropriate training set selected by the PCP. The ANN predicts the short-term traffic flow using e^*. After each estimation process, the training set is renewed for the new traffic situation and the new e^* is determined. These steps are repeated after the current data are entered in the PCP.

The studies on short-term traffic flow prediction started in 1979 for the first time [4]. The published results were studied by different authors on different dates [5, 6]. When the studies are examined according to the used techniques, four different categorizations are used for short-term traffic forecast: naive, parametric, non-parametric and hybrid [6].

In the naive approach, an attempt is made to predict short-term traffic by simple processing of traffic data. Naive approaches are often used in applications since the need for the computational capacity is low. But generally, the results are unsatisfactory. Examples of this approach are: the use of instantaneous values [7-9], forecasts made by the average of the past values [7, 10-12], usage of both instantaneous and past values [13, 14], and the cluster of days with similar traffic patterns [15, 16].

Parametric approaches, on the other hand, are created by determining the parameters that affect the event and by designing them in predetermined forms. These models generally require fewer data than non-parametric models [6]. When the studies forecasting traffic flow rates are examined, it is seen that the traffic flow values, which regard the event as a time series problem, are estimated from the data of the London ring roads using Seasonal Autoregressive Integrated Moving Average (SARIMA) model [17]. Using the same data, the authors predicted the traffic flow rate with 15-minute intervals using SARIMA+ Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models [18]. The researchers developed adaptive SARIMA models in which SARIMA coefficients are adapted to the traffic flow [19]. In addition, they criticized that the parameters of other studies are constant. There are other studies using Autoregressive Integrated Moving Average (ARIMA) model methods [2, 20-24]. Apart from these models, filtering [25], Gaussian Maximum Likelihood [26] and State Space Model [27] can be listed as parametric approaches.

“Non-parametric” term in Non-parametric Approach does not mean that no parameter exists in the model. There are parameters in this approach, but initially, their numbers and attributes are variable and cannot be set at the beginning [6]. These models are data-driven and make traffic flow predictions by using sophisticated algorithms. ANNs are the most commonly used algorithms. For instance, the traffic flow and speed were predicted using ANNs. The datasets consist of measurements taken every minute for 10 days. When different training algorithms of ANNs were compared, the researchers stated that the networks trained with Adaptive Levenberg Marquardt algorithm produced the best results [28]. In their model, they only model weekdays with ANNs. In this case, the validity of the model for weekends is questionable. The Spinning Network (SPN) approach has been developed, inspired by human memory. As a dataset, they converted a one-year traffic dataset they received from the Virginia Department of Transportation into 5-minute sections. They compared the developed approach with the 3-NNB named approaches using the ANN Nearest Neighbourhood algorithm and used the median of the results. Eventually, they argued that the SPN approach, which they developed, gave better results [28]. However, using the first eleven months in the training set, using only December as a test group, and using too many hidden neurons in the ANNs due to the size of the dataset may have affected ANNS prediction performance negatively. Multiple non-parametric methods, linear genetic programming, multilayer perceptron, and fuzzy logic, were used to compare the traffic flow performance predictions [29]. The researchers trained the ANNS network with a 5-day dataset for the year 2012 [30]. In total, 19 input parameters were used. These inputs generally consist of numbers related to vehicle types, variables related to time, speed and the traffic intensity. Different network architectures and transfer functions were tested, and the results were reported to be satisfactory. However, the smallness of the dataset (480 data records), difficulty of accurate measurement from the field and excess amount of input parameters, can be listed as the disadvantages of the study. The k-Nearest Neighbour (k-NN), which is the other approach used, is based on past observations for each prediction and makes predictions with the help of the nearest traffic situation to the current situation. The authors tried to predict the traffic flow rate, speed, and occupancy with k-NN [31]. The researchers who worked with a 3-week, 10-minute traffic datasets reported that k-NN produced less false results than the naive models. The authors predicted the traffic flow with enhanced k-NN [3]. They tested the models with datasets obtained from [2] which consist of a wide variety of regions, and compared the current traffic flow series with other candidate flow series. Then, they compared their method with the four different approaches used in [2] and the Enhanced k-NN approaches, and detected that the proposed model is superior to the other methods. The researchers continue to introduce new artificial intelligence approaches. For instance, a recently developed deep learning approach is an artificial intelligence technique used to describe the graphic model. In another research, the deep learning model was used to estimate the traffic during a football game and on snowy days [32]. The authors concluded that the deep learning model has a low explanatory power. In another study, short-term traffic estimation was made by using network weight matrix method with temporal and spatial inputs [33].

Recently, the studies on short-term traffic forecasting have focused on hybrid approaches. Hybrid methods come up as methods in which more than one approach are used together. The particle swarm optimization was used to optimize ordinary differential equations [34]. It is noteworthy that the time interval in the study is 0.1 s and the positive and negative effects of exceedingly small selected time intervals should be the subject of future work. In addition to these, the following can be listed as hybrid models [35-40]. The k-means algorithm is used by [41] to group the traffic data. The k-means algorithm was used to divide traffic flow data into categories [42]. The researchers developed a hybrid approach called SpAE-LSTM, which uses temporal and spatial features to predict traffic flows [43]. Another hybrid model was designed using SARIMA and seasonal discrete grey model structure. As a result, the researchers stated that the model short-term traffic forecasts are accurate [44]. The researchers proposed a new hybrid model using ARIMA and Wavelet Neural Network together. With this model, they predicted the traffic flow trend. It was stated that the model was more consistent in both stable and fluctuating conditions than the other two models used in the study [45].

Numerous approaches not mentioned here have been used to estimate the important parameters for traffic management systems such as traffic flow rate, speed, and travel time. These studies have been summarized and elaborated in detail [1, 6, 46].

To sum up, it is difficult to identify which of these models work better [3]. Since the data time intervals and performance criteria of the dataset used in these studies are different, it is difficult to confirm this determination. When the studies conducted in the short-term traffic prediction are examined in general, it is understood that ANN and ANN-based approaches give more accurate results [6]. However, the studies are often performed in different model combinations (Hybrid models) or in different versions of ANN-like models. This study expresses an idea about making the dataset more suitable for training a model. This is achieved by dynamically clustering the data in accordance with the current traffic flow. Thus, the model ability to make short-term traffic flow estimation and the consistency of the estimates are increased.

This paper begins with the introduction, followed by the review of the literature, in which relevant work is discussed in detail. Then, the PCP method details are presented. After that, the data used in the test phase and the test results are given. In the last section, discussions and general conclusions about the test results are shared.

Notation

F_i - Forecast value;
N - Number of observations;
O_i - Observed value;
Tr_t - Training set at time number t;

$\bar{C}$ - Vector of the mean value vectors;
$\{E\}$ - All elected set candidate vectors;
$\{F\}$ - Aggregated final vector;
$\{P\}$ - Aggregated preliminary vector;
d_i - Euclidean distance;
k - Number of subsets;
m - Number of traffic flow data in Tr_t;
s - Number of period vectors;
t - Recent time number;
x_t - Recent traffic flow.
x_{t+1} - Future traffic flow.
$\{\bar{c}_i\}$ - The $[e_i]$ column averages vector;
$\{p\}$ - Preliminary period vector;
$\{f\}$ - Final period vector;
$[e_i]$ - Elected set candidate in matrix form;
$[e^*]$ - Elected set;
α - Ratio to determine the k ($k=\alpha \cdot m$);
μ - Column average value.

2. PERIODIC CLUSTERING AND PREDICTION METHOD (PCP)

In this study, the Periodic Clustering and Prediction (PCP) approach which is the cooperation of k-Means Clustering Algorithm and the ANNs has been developed for short-term traffic flow prediction. There are three main stages in this approach. First, data to be used are determined and outlier detection and smoothing are applied to the data. Second, the dataset is divided into periods and the periods are grouped with k-Means Clustering. Then the Elected Set $[e^*]$ is determined and ANN is trained using $[e^*]$. The flowchart of the developed approach is given in *Figure 1*.

It would be useful to explain some parameters used before going through the phases of the PCP. Each square shape in *Figure 2* shows the 15-minute traffic flow. $\{Tr_t\}$ refers to the training data that ANN will use at time number t. The number of traffic flows of $\{Tr_t\}$ is indicated by m and this number is taken as a constant during the prediction process. In addition, x_t indicates the recent traffic flow, while x_{t+1} indicates the predicted flow.

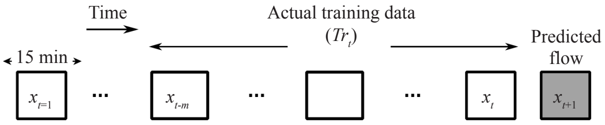

Figure 2 – Training data position on time number t

2.1 Training data and data pre-processing

PCP is a data-driven approach and needs a certain amount of time-series data in order to be able to perform the training and prediction process. For this reason, first $\{Tr_t\}$ is determined from the raw traffic data after checking if there are no missing data and completing the missing data if they exist. The collection of raw traffic flow data can be done with various detectors, receivers or cameras. Especially in long-term counting operations, short-term malfunctions in counting devices and values contrary to

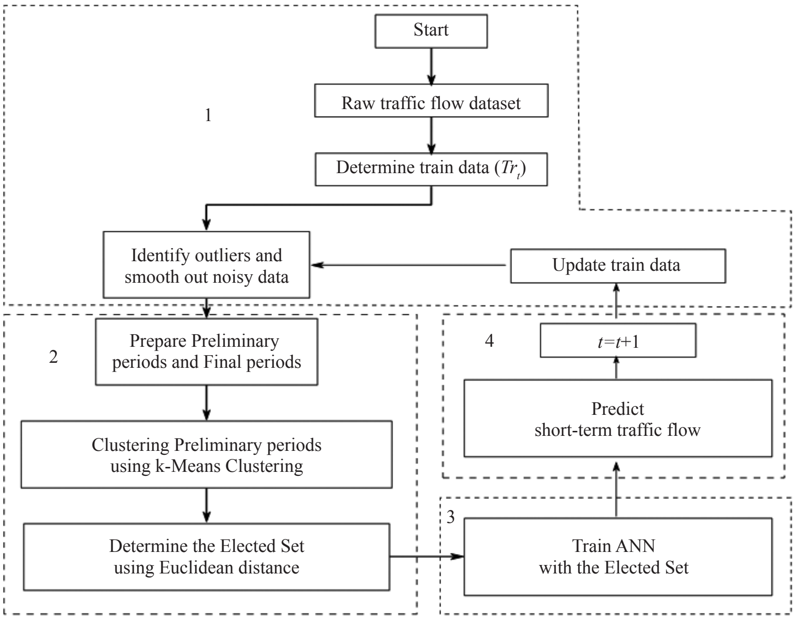
Start
Raw traffic flow dataset
1
Determine train data (Tr_t)
Identify outliers and smooth out noisy data
Update train data
Prepare Preliminary periods and Final periods
2
4
$t=t+1$
Predict short-term traffic flow
Clustering Preliminary periods using k-Means Clustering
3
Determine the Elected Set using Euclidean distance
Train ANN with the Elected Set

Figure 1 – Periodic Clustering and Prediction flow chart

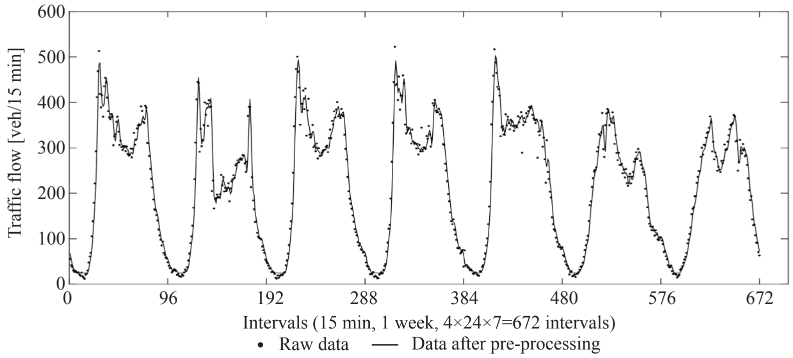

Figure 3 – A sample of smoothing raw data

the general traffic pattern need to be corrected. For this reason; after determination of $\{Tr_t\}$, the outliers are determined, and the training series is smoothed (see *Figure 1*). The values in the traffic flow series likely to reduce the modelling performance were determined using Hampel identifiers [47, 48]. Hampel identifiers are described in the literature as the most effective and efficient outlier detection algorithm [49, 50]. Then, the $\{Tr_t\}$ smoothing process is performed using local regression using weighted linear least squares (loess) [51, 52]. The most important advantage of this method is that it does not require any assumption about the dataset [53]. This pre-processing on traffic flow data is only applied to $\{Tr_t\}$. For this reason, there is no change in the value of x_{t+1} that is supposed to be predicted. Raw data for one week and the related traffic flow rates after pre-processing are shown in *Figure 3*.

2.2 Composing periods and k-Means Clustering

After the data are identified and pre-processed, $\{Tr_t\}$ is separated into two types of periods. These are preliminary period vector $\{p\}$ and final period vector $\{f\}$. These vectors are assembled in the aggregated preliminary vector $\{P\}$ and the aggregated final vector $\{F\}$.

$\{Tr_t\}$ time series vector:

$$Tr_t = \{x_{t-m+1},\ x_{t-m+2},\ \ldots,\ x_t\} \tag{1}$$

Preliminary period vectors:

$$p_1 = \{x_{t-m+1},\ x_{t-m+2},\ x_{t-m+3}\},\ \ldots,\ p_s = \{x_{t-5},\ x_{t-4},\ x_{t-3}\}, \text{ thus, } P = \{p_1,\ p_2,\ \ldots,\ p_s\} \tag{2}$$

Final period vectors:

$$f_1 = \{x_{t-m+4},\ x_{t-m+5},\ x_{t-m+6}\},\ \ldots,\ f_s = \{x_{t-2},\ x_{t-1},\ x_t\}, \text{ thus, } F = \{f_1,\ f_2,\ \ldots,\ f_s\} \tag{3}$$

where s is the number of period vectors.

The k-means algorithm is an unsupervised clustering algorithm used in data mining [54, 55]. This algorithm divides a dataset into similar k subsets. In the developed approach, the k-means algorithm separates the final vectors into subsets according to their similarities. At the same time, PCP places the first vectors in subsets using the indices of the final vectors.

As a result of the clustering process, in a previously determined cluster amount of subsets (k) emerge. Thus, k datasets emerge as candidates for the training of ANNS. These subsets are Elected Set Candidates $[e_i]$ and consist of a combination of similar preliminary period vectors. The $[e_i]$ are arranged in the form of matrices and all candidate vectors $\{E\}$ are expressed as follows:

$$E = \{e_1,\ e_2,\ \ldots,\ e_k\},\ \ (i = 1,\ 2,\ \ldots,\ k) \tag{4}$$

Then, the averages of $[e_i]$ columns are calculated, and these values are expressed in $\{\overline{c}_i\}$ vector format as follows:

$$\overline{c}_i = \{\mu_1,\ \mu_2,\ \ldots,\ \mu_n\},\ \ i = 1,\ 2,\ \ldots,\ k \tag{5}$$

Thus, what arises from the column averages of each mean value,

$$\overline{C} = \{\overline{c}_1,\ \overline{c}_2,\ \ldots,\ \overline{c}_k\} \tag{6}$$

occurs in this form; where:

$\overline{c}_i$ - vector of mean values of i, (i=1, 2, ..., k);

μ - column average value;

n - number of columns for $[e_i]$ (n=3 was accepted for this study);

$\overline{C}$ - vector of the mean value vectors.

After the operations described above, $\{e^*\}$ can be determined. For this, the Euclidean distance (d_i) between $\{f_s\}$ and $\{\overline{c}_i\}$ vector is calculated with *Equation 7*.

$$d_i = \sqrt{\sum_{i=1}^{k} [\{f_s\} - \{\overline{c}_i\}]^2} \tag{7}$$

As a result, the set giving the shortest Euclidean distance is selected and used for the training of ANN. This process is repeated every $t+1$ and the $\{e^*\}$ is re-determined according to the new traffic flow situation.

2.3 Artificial Neural Network and traffic flow prediction

ANN is an artificial intelligence technique developed by being influenced by the biological neurons and their connections. Nowadays, ANNs are used successfully in areas such as prediction, image processing, clustering, etc. They are also used for traffic flow prediction, as discussed in Chapter 2. The ANN used in the PCP approach is trained with the Levenberg-Marquardt Algorithm [56, 57]. The number of neurons in the hidden layer (*Nhn*) has a significant effect on the performance. For this reason, *Nhn*, giving the best result according to the test results, needs to be determined.

Traffic flow prediction can be done after $\{e^*\}$ is determined as the training set and the ANN is trained.

To compare the result of the prediction with the actual values and to measure the performance of the PCP after the prediction, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) given in *Equations 8-10* were taken as performance criteria.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|F_i - O_i| \quad (8)$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{F_i - O_i}{O_i}\right| \cdot 100 \quad (9)$$

$$RMSE = \left[\frac{\sum_{i=1}^{N}(F_i - O_i)^2}{N}\right]^{\frac{1}{2}} \quad (10)$$

where:
F_i - i-th forecast value;
O_i - i-th observed value;
N - number of observations.

These performance measures are frequently used in the literature, and more information about these performance measures can be obtained by the reader from resources [58-60].

In order to better understand the performance of the approach at different traffic flow rates, the traffic flow rates were divided into groups as in *Table 1*.

3. EXPERIMENTS AND RESULTS

The fact that the dataset is of a size that extends to a year and is collected from different regions, facilitates to model and test these models reliably.

Table 1 – Traffic volume groups

Volume groups	Group description
Group 1 (G1)	$0 \le x < 500$
Group 2 (G2)	$500 \le x < 1{,}000$
Group 3 (G3)	$1{,}000 \le x < 1{,}500$
Group 4 (G4)	$1{,}500 \le x < 2{,}000$
Group 5 (G5)	$2{,}000 \le x$

Note: x = vehicle number per hour per lane

In this study, the same dataset is used by [2] and then by [3]. The researchers used simple screening procedures, i.e. threshold test and hang-on test, to eliminate the erroneous data. They completed the missing data in the dataset using SARIMA (1,01) $(0,1,1)_{672}$ model. This dataset is collected from 32 different stations located on freeways and motorways of the United States and the United Kingdom. The datasets have been collected at 15-minute intervals and collection intervals vary from 3 months to 12 months. The missing data in the dataset were completed using the SARIMA (1,01) $(0,1,1)_{672}$ model. For further information see [2, 24].

3.1 Effect of the *k*, *m* and *Nhn* parameters on performance

There are three main parameters that can affect the traffic flow prediction performance. These are the length of the main training series (*m*), the number of the subsets belonging to the main training series divided by the k-means algorithm (*k*), and the number of hidden neurons belonging to ANN in the system (*Nhn*). The effects of these parameters on performance for forecasting the number of vehicles passing within the 15-minute time slot are given in *Table 2*.

Three different values of *m* were examined to observe their effects on the performance, which are 7 days (7×24×4=>*m*=672), 14 days (14×24×4=> *m*=1,344) and 30 days (30×24×4=>*m*=2,880). Moreover, for each value of *m*, different *Nhn* values shown in *Table 2* and different *α* values, which are the ratios used to determine the *k* value, were analysed. For example, in a monthly training set with *m*=2,880, the *k* value for *α*=0.02 is 2,880×0.02≈58. Thus, the aim is to make it easier to compare the training sets at varied sizes. MAE, MAPE, and RMSE given in *Equations 2-4* were used as performance measures.

Table 2 – Average prediction errors for the parameters Nhn, m, and α [veh/15 min/ln]

Length of *m*	*Nhn*	*α = k/m*																	
		0.005			0.01			0.02			0.03			0.05			0.1		
		MAE	*MAPE*	*RMSE*	*MAE*	*MAPE*	*RMSE*	*MAE*	*MAPE*	*RMSE*	*MAE*	*MAPE*	*RMSE*	*MAE*	*MAPE*	*RMSE*	*MAE*	*MAPE*	*RMSE*
1 week	5	15.64	8.58	24.95	16.14	8.26	27.26	16.39	8.31	33.82	16.63	8.57	27.93	16.83	8.75	27.87	16.44	8.85	26.20
	7	15.43	8.37	24.86	16.31	8.31	30.10	16.27	8.35	27.00	16.62	8.49	27.78	17.07	8.70	27.83	16.78	8.97	27.06
	9	15.29	8.17	24.77	16.15	8.29	26.58	16.58	8.52	27.84	17.61	8.95	31.46	17.00	8.78	27.41	18.18	9.83	29.54
	11	15.37	8.20	24.91	16.30	8.31	27.45	16.87	8.64	27.64	17.51	8.97	29.40	17.92	9.16	30.21	16.72	8.76	27.11
	13	15.42	8.31	25.02	16.35	8.31	27.69	17.26	8.82	29.36	18.01	9.19	31.09	19.28	10.16	31.91	18.63	10.32	30.41
	15	15.42	8.22	25.17	16.18	8.28	26.58	18.08	9.17	33.40	17.74	9.24	29.72	18.13	9.40	30.05	17.73	9.55	28.20
2 weeks	5	15.95	8.09	26.07	15.82	8.15	27.59	15.81	8.06	26.48	16.11	8.28	32.43	15.64	8.49	27.08	16.57	9.02	26.91
	7	15.82	8.10	25.80	15.89	8.12	26.54	15.32	7.91	25.06	15.81	8.07	25.92	15.65	8.32	25.65	17.04	9.47	27.64
	9	15.92	8.15	25.80	15.79	8.15	25.69	15.87	8.11	26.73	16.60	8.39	29.39	16.58	8.94	27.48	15.84	9.00	25.67
	11	15.97	8.16	26.24	15.94	8.14	26.63	16.03	8.26	26.57	16.45	8.44	28.24	15.88	8.55	25.99	16.26	9.23	26.07
	13	16.23	8.22	27.88	16.20	8.24	27.14	15.94	8.13	26.62	16.60	8.33	28.44	16.75	9.04	27.24	16.45	9.54	26.41
	15	16.04	8.13	26.79	16.42	8.33	27.88	16.38	8.32	28.22	16.52	8.46	27.75	16.48	8.95	26.67	16.61	9.05	26.92
1 month	5	15.57	8.05	27.24	14.96	7.69	24.55	14.82	7.60	24.55	15.10	7.96	25.15	15.46	8.44	25.31	15.82	8.64	25.90
	7	15.38	7.93	25.15	15.23	7.80	25.89	14.92	7.65	24.56	15.05	7.91	24.79	15.94	8.58	26.40	16.30	8.78	27.00
	9	15.66	8.12	25.93	15.00	7.67	24.75	15.07	7.71	24.77	15.28	8.13	25.23	15.79	8.37	25.81	16.16	8.56	26.41
	11	15.44	7.93	25.45	15.09	7.69	24.93	15.51	7.99	26.73	15.19	8.13	25.10	16.08	8.75	26.43	16.45	8.95	27.04
	13	15.55	8.05	25.25	15.21	7.78	25.16	15.28	7.93	25.38	15.75	8.42	26.16	16.22	8.67	26.94	16.59	8.87	27.56
	15	15.44	7.97	25.29	15.30	7.77	25.41	15.80	8.26	27.47	15.99	8.70	26.74	16.46	8.97	27.54	16.84	9.17	28.17

When the error values in *Table 2* were examined, it was observed that the best result was the combined length of one month of the main training set with α of 0.02 and *Nhn* of 5. When the error values of α=0.02 are examined, it is seen that as *m* increases the error values decrease. Again, looking at the value of *m* in the one-month length where the best result is produced, the performance is improved by the increase of the α value up to 0.02. However, it is understood from the one-month lines of *Table 2* that further increase of alpha value leads to an increase in error values.

It is necessary to determine how the clustering improves the prediction performance. Therefore, re-prediction was performed for *k* as 1, training set length as 1 month, and the *Nhn* as 5. As a result, it was calculated that the MAE value of 14.8, which belongs to PCP working with the improved training set, was increased to 17.46 veh/15 min per lane, the MAPE value of 7.6% was increased to 9.73% and the RMSE value of 24.55 was increased to 26.07 veh/15min/ln.

As a result of testing with different variants of *k*, *m*, *Nhn* numbers; it was observed that as the training set gets larger, the error values decrease. However, the results of the growth of *m* values over one month have remained unclear within this study limits, since the *m* analyses have been tested for a maximum of one month. It would be appropriate to examine the increase in these values in further studies. The *k* value of the k-means algorithm is determined by the α ratio. It has been observed that the number of clusters up to 2% of the number of data included in the training set affects the performance in a good way. It is also calculated that because of using *k*=1, which means the whole training set is used for ANN training without clustering, the MAPE value increases by about 30%. It is understood that the clustering process prior to the training in the light of these results may improve the performance. Moreover, the magnitudes of the periods are kept constant in this study, and the performance of different large periods is expected to have a potential to be examined in further studies. In addition, the performance of different smoothing and outlier methods can be examined in future studies.

Matlab 2016a version was used in the execution of PCP. After starting the PCP with 500 iterations, the average computational time required for one iteration was found to be 0.3284 seconds. Considering that short-term estimates are made at 15-minute intervals, this calculation time is quite satisfactory.

3.2 Performance comparisons belonging to traffic groups and intraday hours

In this section, firstly, the results of the performance of PCP in different traffic volumes and hours during the day are discussed. Then, PCP was compared with five different prediction methods used for short-term traffic prediction in previous studies. These methods are: Enhanced k-NN, EXPRW, BATCH, KF and AKF. To sum up these methods briefly, the Enhanced k-NN is a method looking for similar patterns in order to forecast the traffic flow [3]. The other methods, namely EXPRW, BATCH, KF and AKF are used by [2]. In short, EXPW uses seasonal exponential smoothing to capture traffic flow profiles. The BATCH method makes future estimates by generating SARIMA and Autoregressive Conditional Heteroskedasticity (GARCH) models. The KF method was a standard Kalman filter and uses seasonal exponential smoothing method. Finally, AKF, unlike the KF method, consists of an adaptive filter.

Revealing the response of a prediction model to different conditions is important to identify the strengths and weaknesses of the PCP. Thus, the performance of PCP in different traffic conditions, i.e. where the traffic flow value takes different values, is illustrated in *Figures 4* and *5*. After prediction and error evaluation for each station, *Figures 4* and *5* are illustrated with the help of 36 performance criterion values of the stations.

The PCP performance determined for the traffic groups expressed by the ranges of traffic flow values is presented in *Figure 4*. The traffic groups are explained in *Table 1* depending on the traffic volume. Thus, the accuracy of PCP predictions could be discussed due to the change of traffic from low volumes to high volumes. In predictions, low-volume traffic flows, even small scalar errors can reach large percentages. Therefore, in *Figure 4*, MAPE values of G1 were higher than MAPE values of other traffic groups. In the box diagram, it is possible to see the upper and lower error values of the traffic forecasts of 36 stations from the whiskers of the diagram. For example, in one of the stations, the MAPE value of G1 decreased to 4%. At another station, however, this value increased to 8%. Other station errors changed between these two values. From the box diagram, it is also possible to read the percentage of the stations in which the errors range. The portion between the top and bottom whiskers and the box represents 25% of the data used.

If the results for the groups are investigated with the help of *Figure 4*, it is understood that the MAPE value falls as we go from G1 to G5. However, RMSE and MAE values seem to increase slowly. When the whiskers of MAPE plot value are examined through *Figure 4*; it is observed that the difference is about 4% for G1, decreases abruptly in other groups, and decreases even to 1.5% for G5. It is understood that the average MAPE value of these five groups is 3.83%. The values of MAE and RMSE increase with the increase of traffic flow value. The mean MAE and RMSE values for G1 were found to be 13.12 and 16.95 veh/h/ln, respectively. The highest error values were observed in G5 as expected and were determined to be 50.80 and 65.76 veh/h/ln, respectively.

The error values during the day were analysed on an hourly basis and the results are shown in *Figure 4*. When the average MAPE values are examined, it is seen that the error values decrease to 2% between 1 p.m. and 5 p.m. It is detected that it is around 4%, at the peak traffic hours in the morning. The mean values of MAE and RMSE values were monitored

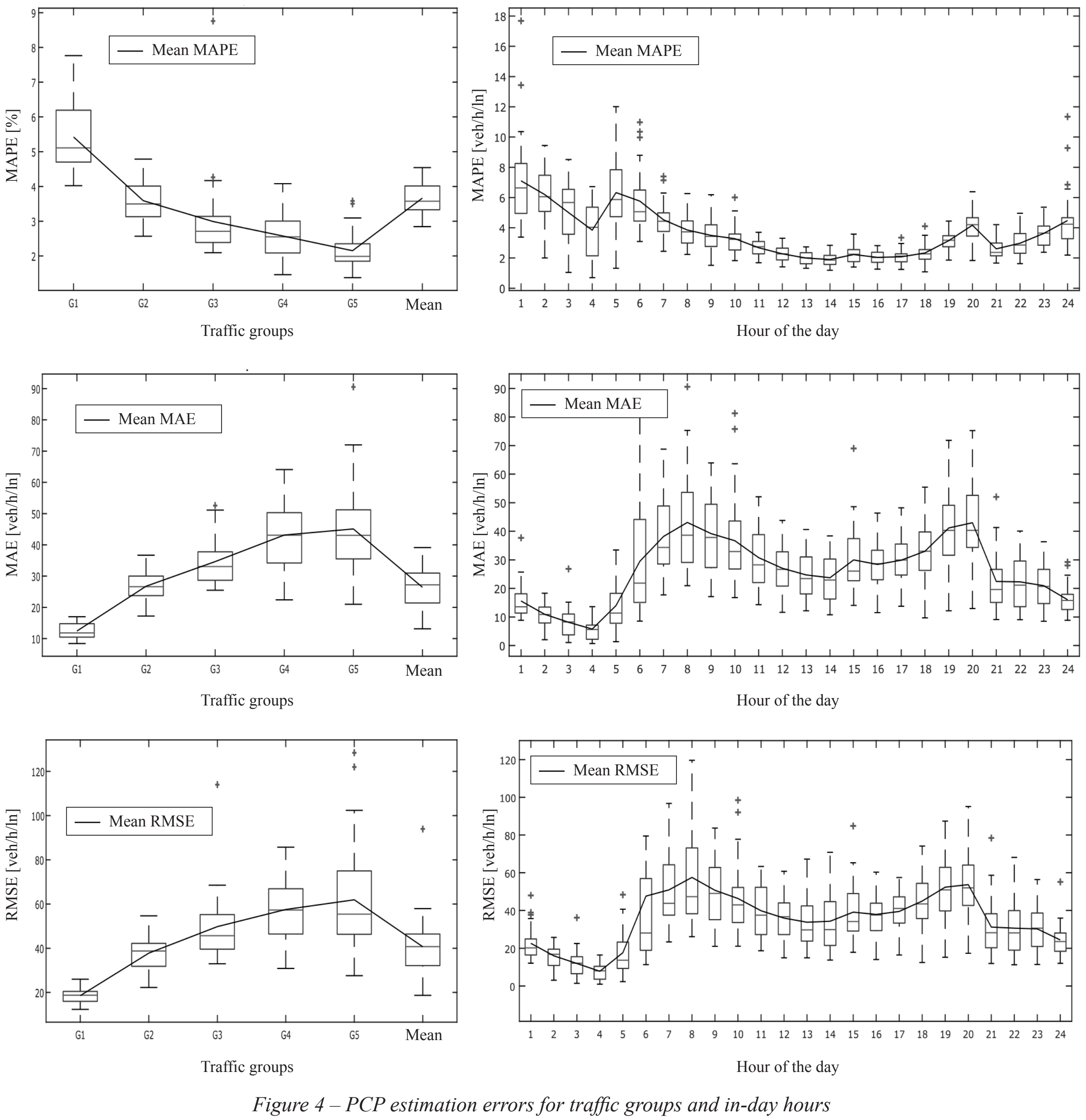

Figure 4 – PCP estimation errors for traffic groups and in-day hours

as 45 and 55 veh/h/ln, respectively. The lowest error values exist generally around 4 a.m. at night when the lowest traffic flow is observed.

The change of R^2 values calculated according to the estimation results of PCP according to traffic groups and intraday hours is illustrated in *Figure 5*. According to traffic groups, higher R^2 values were more frequent in cases representing low traffic flow values. With the increase of traffic flow value (from G1 to G5) R^2 decreased. In G5 representing very high traffic flow values, it is seen from *Figure 5* that the predictive performance of PCP decreased dramatically. However, this is not exactly true, since in some stations the number of samples in the G5 case was very small. Therefore, the calculated R^2 for some stations was too low. However, if the outliers were not considered, it is clearly seen from *Figure 4* that the average R^2 for G5 was 70%. The average R^2 value of all traffic groups was found to be the lowest 80% and the highest 95%, respectively. These R^2 values show that the predicted traffic flow of the developed system is statistically highly accurate.

The mean error values of the developed PCP method are compared in *Figure 6* with the Enhanced k-NN, EXPRW, BATCH, KF and AKF methods. When the mean error values, plotted against the traffic groups in *Figure 6*, are examined; it is understood that the MAPE value is generally decreasing towards G1 to G5. MAE and RMSE values, on the other hand, are usually increasing in all approaches. It is observed that the error values of PCP are lower than the other approaches with one exception (*Figure 5*). This exception occurred at the point where the Enhanced k-NN method had an error of about 1% less than the PCP for the MAPE value of G1.

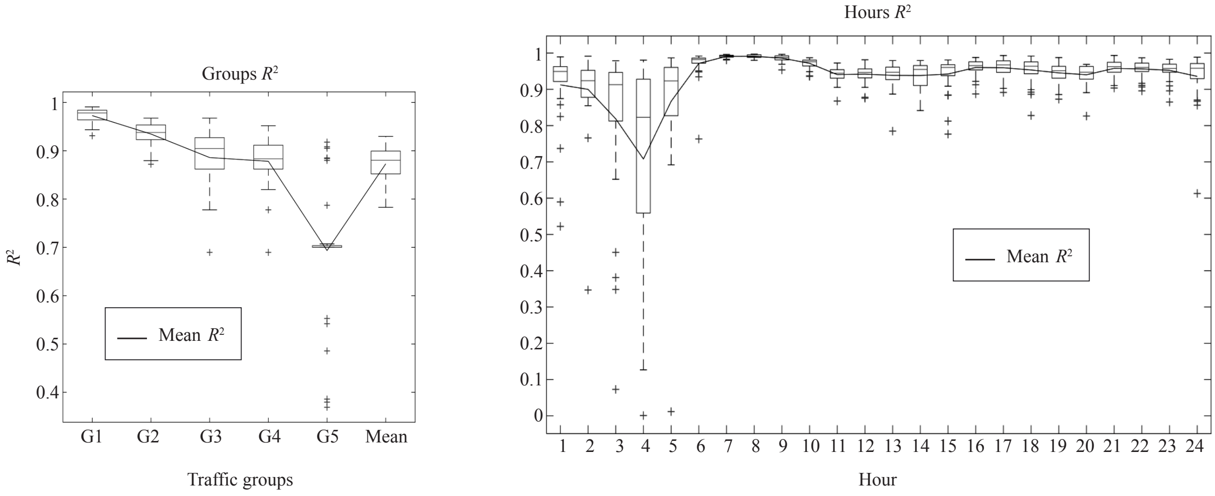

Figure 5 – R^2 values of PCP predictions for traffic groups and in-day hours

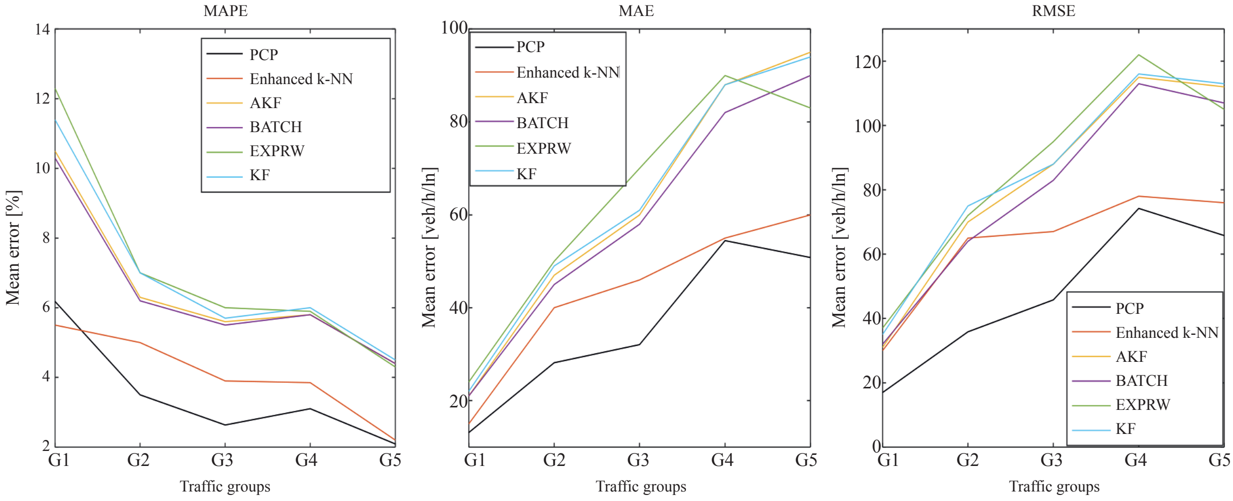

Figure 6 – Error comparisons of PCP and other models

When the error values given in *Figure 5* are analysed in more detail, the approaches other than PCP and Enhanced k-NN produced more than 10% MAPE in G1. The PCP and Enhanced k-NN produced less than 10 MAEs for the same traffic group, while the other approaches produced over 20 MAEs. The PCP and Enhanced k-NN produced less than 10 MAEs for the same traffic group, while the other approaches produced 20 veh/h. In RMSE, PCP achieved higher performance compared to the other approaches. It is observed that the MAPE values of Enhanced k-NN in G4 and G5 groups with the highest traffic flow rate were 3.8% and 2.3%, and PCP is 3.1% and 2.1%, respectively. It is understood from *Figure 6* that the greatest difference is realized in G3, which is the group observed from many points in daytime hours during the day. This difference occurred at 14 veh/h/ln for MAE and 21 veh/h/ln for RMSE. In the light of the above comparisons, the developed PCP method seems to perform better than the other referenced methods.

In brief, the following evaluations can be made according to the results discussed in this section. The MAPE value is encountered as high, especially in case of small traffic flow values. For this reason, it is understood that the high error percentages made even for a few points increase the average MAPE value. The mean MAPE of the PCP being higher than the Enhanced k-NN error values for G1 indicates that the PCP makes mistakes at fewer points with high percentages, and the Enhanced k-NN makes mistakes at more points with small percentages. This is also evident from the analysis of the mean values of MAE and RMSE. The PCP method yielded better results compared to the other approaches examined at all points, including G1. The G1 state indicates the lowest traffic flow condition and usually occurs at midnight hours. In terms of traffic management, these cases do not have critical importance since the road capacity is already more than the existing traffic. However, for groups with a high and remarkably high traffic flow rate (G3-G5), the short-term traffic flow rate prediction becomes more important. For all these values, it was determined that PCP produces lower error values for all performance criteria.

4. CONCLUSION

The amount of road traffic flow increases and becomes difficult to control over time. Nowadays, sophisticated systems have started to be used to control this situation. Models that can make accurate traffic estimation have the potential to make these systems more efficient. For this purpose, studies for developing such models are in progress. These models are generally developed based on data. Therefore, the main objective of this study was to develop an algorithm that finds more suitable datasets for these data-driven models. In addition, the performance of this approach has been tested in short-term traffic flow forecasting.

To evaluate the performance of PCP, comparisons were made with previous models and this evaluation has shown that the PCP method makes fewer errors in predicting traffic flows than the Enhanced k-NN, EXRW, BATCH, KF and AKF methods. The results of this research support the idea that new and more accurate approaches to predict short-term traffic flow can be developed by the researchers.

More research is required to determine the efficacy of PCP. For example, the dataset used contains only uninterrupted traffic flow data. The effect of PCP under interrupted flow can be investigated in further studies. In addition, the performance of PCP was not investigated against the missing values in the data. This limitation should be clarified in the future.

PCP introduced a new periodic clustering approach to obtain more appropriate training data for ANNs. The developed PCP approach only needs a small amount of past traffic flow data. In addition, k-means and ANNs algorithms used in today's computers can work fast. For these reasons, the developed method can be used in traffic applications. In addition, the usage area of PCP is not limited to estimating only the traffic flow, but the idea is to use it successfully in the time series.

ACKNOWLEDGEMENT

The author would like to thank the Minnesota Department of Transportation, the Washington State Department of Transportation, the Maryland Department of Transportation, and the United Kingdom Highways Agency for providing the data used in this study. Also, the author would like to thank Dr. Billy Williams and Dr. Jianhua Guo for providing these data.

ERDEM DOĞAN, Ph.D.
E-mail: edogan@kku.edu.tr
Kırıkkale Üniversitesi, Mühendislik Fakültesi
İnşaat Mühendisliği Bölümü
Kırıkkale, Yahşihan, 71451, Türkiye

PERİYODİK KÜMELEME VE SEÇİLMİŞ KÜMELER KULLANAN YAPAY ZEKA YAKLAŞIMI İLE KISA VADELİ TRAFİK AKIŞ TAHMİNİ

ÖZET

Trafik akışındaki dalgalanmaların rastgelelik içermesi, kısa dönemli trafik akış tahminini isabetli yapabilen bir yaklaşımın geliştirilmesini zor hale getirmektedir. Bu nedenle araştırmacılar, isabetli kısa vadeli trafik tahmini yapabilen yeni yaklaşımları geliştirmek için çalışmalar yapmaya devam etmektedirler. Bu çalışmada, bir tahmin modelinin daha isabetli tahminler yapabilmesi için eğitim setini dinamik güncelleyen Periyodik Kümeleme ile Tahmin (PCP) isimli yeni bir algoritma geliştirilmiştir. PCP, seçilmiş kümeler ve periyodik kümeleme yaparak bir modelin eğitim setini özelleştiren bir algoritmadır. Bu çalışmada, PCP, Yapay Sinir Ağlarının (YSA) eğitim setini özelleştirmek ve YSA'nın trafik tahmin performansını iyileştirmek için kullanılmıştır. PCP' nin YSA performansını artırma yeteneğinin düzeyini belirlemek için, A.B.D. ve Birleşik Krallık otoyollarına ait çok miktarda trafik akış verisi kullanılmış ve analizler yapılmıştır. Analizlerde, önerilen yaklaşımın sağlamlığı, literatürde yaygın kullanılan performans kriterleri ile belirlenmiştir. Sonuçlar incelendiğinde, PCP' nin ortalama tahmin hatalarının, diğer yaklaşımların hata değerlerinden daha düşük olduğu görülmüştür. Ayrıca, PCP tahminlerine ait yüzdelik hatalarının, trafik akış değerlerinin artması ile azaldığı da anlaşılmıştır. Elde edilen olumlu sonuçlar, PCP' nin ANN modelinin tahmin performansını arttırdığını ve gerçek zamanlı trafik kontrol sistemlerinde kullanılabilir olduğu ve göstermiştir.

ANAHTAR KELİMELER

trafik tahmini; eğitim seti; kısa vadeli tahmin;
k-ortalamaları; yapay sinir ağları;

REFERENCES

[1] Vlahogianni EI, Karlaftis MG, Golias JC. Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies*. 2014;43: 3-19. Available from: doi:10.1016/j.trc.2014.01.005

[2] 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. Available from: doi:10.1016/j.trc.2014.02.006

[3] 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. Available from: doi:10.1016/j.trc.2015.08.017

[4] Ahmed MS, Cook AR. Analysis of Freeway Traffic Time-Series Data by Using Box-Jenkins Techniques. *Transportation Research Record*. 1979. Available from: doi:10.3141/2024-03

[5] Vlahogianni EI, Golias JC, Karlaftis MG. Short-term traffic forecasting: Overview of objectives and methods. *Transport Reviews*. 2004;24(5): 533-557. Available from: doi:10.1080/0144164042000195072

[6] Van Lint H, van Hinsbergen C. Short-term traffic and travel time prediction models. *Artificial Intelligence Applications to Critical Transportation Issues*; 2012. p. 22-41.

[7] Nikovski D, Nishiuma N, Goto Y, Kumazawa H. Univariate short-term prediction of road travel times. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*; 2005. p. 1074-1079. Available from: doi:10.1109/ITSC.2005.1520200

[8] Huisken G, van Berkum EC. A Comparative Analysis of Short-Range Travel Time Prediction Methods. *TRB 2003 Annual Meeting CD-ROM*. 2003; 21 p.

[9] Wu CH, Ho JM, Lee DT. Travel-time prediction with support vector regression. *IEEE Transactions on Intelligent Transportation Systems*. 2004. p. 276-281. Available from: doi:10.1109/TITS.2004.837813

[10] Park D, Rilett LR. Forecasting multiple-period freeway link travel times using modular neural networks. *Transportation Research Record: Journal of the Transportation Research Board*. 1998;1617: 163-170.

[11] Kamarianakis Y, Prastacos P. Forecasting Traffic Flow Conditions in an Urban Network: Comparison of Multivariate and Univariate Approaches. *Transportation Research Record: Journal of the Transportation Research Board.* 2003;1857(1): 74-84. Available from: doi:10.3141/1857-09

[12] Eglese R, Maden W, Slater A. A Road TimetableTM to aid vehicle routing and scheduling. *Computers and Operations Research*. 2006;33(12): 3508-3519. Available from: doi:10.1016/j.cor.2005.03.029

[13] Hobeika AG, Kim C. Traffic-Flow-Prediction Systems Based on Upstream Traffic a.G. *Vehicle Navigation and Information Systems Conference, 1994. Proceedings*. IEEE; 1990. p. 345-350.

[14] Smith BL, Williams BM, Keith Oswald R. Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C: Emerging Technologies*. 2002;10(4): 303-321. Available from: doi:10.1016/S0968-090X(02)00009-8

[15] Chung E. Classification of traffic pattern. *Proc. of the 11th World Congress on ITS*; 2003. p. 4-6.

[16] Wild D. Short-term forecasting based on a transformation and classification of traffic volume time series. *International Journal of Forecasting.* 1997;13(1): 63-72. Available from: doi:10.1016/S0169-2070(96)00701-7

[17] 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.

[18] Guo J, Williams BM, Smith BL. Data Collection Time Intervals for Stochastic Short-Term Traffic Flow Forecasting. *Transportation Research Record: Journal of the Transportation Research Board.* 2008;2024(1): 18-26. Available from: doi:10.3141/2024-03

[19] Shekhar S, Williams B. Adaptive seasonal time series models for forecasting short-term traffic flow. *Transportation Research Record: Journal of the Transportation*

Research Board. 2008;(2024): 116-125.

[20] Zeng D, Xu J, Gu J, Liu L, Xu G. Short term traffic flow prediction using hybrid ARIMA and ANN models. *Proceedings - 2008 Workshop on Power Electronics and Intelligent Transportation System, PEITS 2008*; 2008. Available from: doi:10.1109/PEITS.2008.135

[21] Lin SL, Huang HQ, Zhu DQ, Wang TZ. The application of space-time arima model on traffic flow forecasting. *Proceedings of the 2009 International Conference on Machine Learning and Cybernetics*. IEEE; 2009. p. 3408-3412. Available from: doi:10.1109/ICMLC.2009.5212785

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

[23] Zhou B, He D, Sun Z. Traffic predictability based on ARIMA/GARCH model. *Next Generation Internet Design and Engineering, 2006. NGI'06. 2006 2nd Conference on*. IEEE; 2006. 8 p.

[24] Guo J, Huang W, Williams BM. Integrated Heteroscedasticity Test for Vehicular Traffic Condition Series. *Journal of Transportation Engineering*. 2012;138(9): 1161-1170. Available from: doi:10.1061/(ASCE)TE.1943-5436.0000420

[25] Yang J-S. Travel time prediction using the GPS test vehicle and Kalman filtering techniques. *Proceedings of the 2005, American Control Conference, 2005*; 2005. p. 2128-2133. Available from: doi:10.1109/ACC.2005.1470285

[26] Lin W-H. A Gaussian maximum likelihood formulation for short-term forecasting of traffic flow. *ITSC 2001. 2001 IEEE Intelligent Transportation Systems. Proceedings (Cat. No.01TH8585)*. 2001; p. 150-155. Available from: doi:10.1109/ITSC.2001.948646

[27] Stathopoulos A, Karlaftis MG. A multivariate state space approach for urban traffic flow modeling and prediction. *Transportation Research Part C: Emerging Technologies*. 2003;11(2): 121-135. Available from: doi:10.1016/S0968-090X(03)00004-4

[28] Dunne S, Ghosh B. Regime-Based Short-Term Multivariate Traffic Condition Forecasting Algorithm. *Journal of Transportation Engineering*. 2012;138(4): 455-466. Available from: doi:10.1061/(ASCE)TE.1943-5436.0000337

[29] Zargari SA, Siabil SZ, Alavi AH, Gandomi AH. A computational intelligence-based approach for short-term traffic flow prediction. *Expert Systems*. 2012;29(2): 124-142. Available from: doi:10.1111/j.1468-0394.2010.00567.x

[30] Kumar K, Parida M, Katiyar VK. Short term traffic flow prediction in heterogeneous condition using artificial neural network. *Transport*. 2015;30(4): 397-405. Available from: doi:10.3846/16484142.2013.818057

[31] Clark S. Traffic Prediction Using Multivariate Nonparametric Regression. *Journal of Transportation Engineering*. 2003;129(2): 161-168. Available from: doi:10.1061/(ASCE)0733-947X(2003)129:2(161)

[32] Polson NG, Sokolov VO. Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*. 2017;79: 1-17. Available from: doi:https://doi.org/10.1016/j.trc.2017.02.024

[33] Ermagun A, Levinson D. Spatiotemporal short-term traffic forecasting using the network weight matrix and systematic detrending. *Transportation Research Part C: Emerging Technologies*. 2019;104: 38-52.

[34] Chen Y, Yang B, Meng Q, Zhao Y, Abraham A. Time-series forecasting using a system of ordinary differential equations. *Information Sciences*. 2011;181(1): 106-114. Available from: doi:10.1016/j.ins.2010.09.006

[35] Hong WC. Traffic flow forecasting by seasonal SVR with chaotic simulated annealing algorithm. *Neurocomputing*. 2011;74(12-13): 2096-2107. Available from: doi:10.1016/j.neucom.2010.12.032

[36] Hong WC, Dong Y, Zheng F, Lai CY. Forecasting urban traffic flow by SVR with continuous ACO. *Applied Mathematical Modelling*. 2011;35(3): 1282-1291. Available from: doi:10.1016/j.apm.2010.09.005

[37] Hong WC, Dong Y, Zheng F, Wei SY. Hybrid evolutionary algorithms in a SVR traffic flow forecasting model. *Applied Mathematics and Computation*. 2011;217(15): 6733-6747. Available from: doi:10.1016/j.amc.2011.01.073

[38] Abdi J, Moshiri B, Abdulhai B, Sedigh AK. Forecasting of short-term traffic-flow based on improved neurofuzzy models via emotional temporal difference learning algorithm. *Engineering Applications of Artificial Intelligence*. 2012;25(5): 1022-1042. Available from: doi:10.1016/j.engappai.2011.09.011

[39] 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. Available from: doi:10.1016/j.trc.2013.11.011

[40] Feng X, Ling X, Zheng H, Chen Z, Xu Y. Adaptive Multi-Kernel SVM With Spatial-Temporal Correlation for Short-Term Traffic Flow Prediction. *IEEE Transactions on Intelligent Transportation Systems*. 2018; p. 1-13. Available from: doi:10.1109/TITS.2018.2854913

[41] Vlahogianni EI, Karlaftis MG, Golias JC, Kourbelis ND. Pattern-Based Short-Term Urban Traffic Predictor. 2006 *IEEE Intelligent Transportation Systems Conference*. 2006; p. 389-393. Available from: doi:10.1109/ITSC.2006.1706772

[42] Yuan Z, Zhang W, Yang M. A Short-term Traffic Flow Prediction Approach of Neural Network Based on Cluster Analysis. *DEStech Transactions on Engineering and Technology Research*. 2016;(iceta).

[43] Lin F, Xu Y, Yang Y, Ma H. A Spatial-Temporal Hybrid Model for Short-Term Traffic Prediction. *Mathematical Problems in Engineering*. Hindawi; 2019; Article ID 4858546. 12 p.

[44] Song Z, Guo Y, Wu Y, Ma J. Short-term traffic speed prediction under different data collection time intervals using a SARIMA-SDGM hybrid prediction model. *PloS one*. 2019;14(6): e0218626.

[45] Hou Q, Leng J, Ma G, Liu W, Cheng Y. An adaptive hybrid model for short-term urban traffic flow prediction. *Physica A: Statistical Mechanics and its Applications*. 2019;527: 121065.

[46] Desch CH. Conservation of natural resources. *Nature*. 1941;148(3758): 547-549. Available from: doi:10.1038/148547a0

[47] Hampel FR. A General Qualitative Definition of Robustness. *The Annals of Mathematical Statistics*.

1971;42(6): 1887-1896. Available from: doi:10.1214/aoms/1177693054

[48] Hampel FR. The influence curve and its role in robust estimation. *Journal of the American Statistical Association.* 1974;69(346): 383-393. Available from: doi:10.1080/01621459.1974.10482962

[49] Davies L, Gather U. The identification of multiple outliers. *Journal of the American Statistical Association.* 1993;88(423): 782-792. Available from: doi:10.1080/01621459.1993.10476339

[50] Pearson RK. Outliers in process modeling and identification. *IEEE Transactions on Control Systems Technology.* 2002;10(1): 55-63. Available from: doi:10.1109/87.974338

[51] Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association.* 1979;74(368): 829-836. Available from: doi:10.1080/01621459.1979.10481038

[52] Cleveland WS. Lowess: A program for smoothing scatterplots by robust locally weighted regression. *American Statistician*. 1981;35(1): 54-55. Available from: doi:10.1080/00031305.1981.10479306_3

[53] Cleveland WS, Devlin SJ. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association.* 1988;83(403): 596-610. Available from: doi:10.1080/01621459.1988.10478639

[54] MacQueen J. Some Methods for classification and Analysis of Multivariate Observations. *5th Berkeley Symposium on Mathematical Statistics and Probability 1967*; 1967. p. 281-297. Available from: doi:citeulike-article-id:6083430

[55] Forgy EW. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*. 1965;21(3): 768-769. Available from: doi:10.1007/s00442-008-1028-8

[56] Yin X, Zhang J, Wang X. Sequential injection analysis system for the determination of arsenic by hydride generation atomic absorption spectrometry. *Fenxi Huaxue*. 2004;32(10): 1365-1367. Available from: doi:10.1017/CBO9781107415324.004

[57] Marquardt DW. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics.* 1963;11(2): 431-441. Available from: doi:10.1137/0111030

[58] Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*. 2005;30(1): 79-82.

[59] Willmott CJ. Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society.* 1982;63(11): 1309-1313.

[60] Hyndman RJ. Another look at measures of forecast accuracy for intermittent demand. *Foresight: the International Journal of Applied Forecasting.* 2006;4(4): 43-46.