Sigma J Eng & Nat Sci 9 (4), 2018, 405-414

Publications Prepared for Transist 2017,
10th International İstanbul Transport Congress

Research Article

SHORT TERM TRAFFIC FLOW FORECASTING USING ARTIFICIAL NEURAL NETWORKS

Zeynep Idil ERZURUM CICEK*[1], Zehra KAMISLI OZTURK[2]

[1]*Eskisehir Technical University, Dept. of Industrial Engineering, ESKISEHIR;* ORCID:0000-0001-9641-8935
[2]*Eskisehir Technical University, Dept. of Industrial Engineering, ESKISEHIR;* ORCID:0000-0003-3156-6464

Received: 08.10.2018 Revised: 13.12.2018 Accepted: 21.12.2018

ABSTRACT

Traffic flow forecasting is a critical issue in detection of the traffic congestions. Better forecasts provide better routes, less travel time and less traffic bottlenecks. In this study, an existing traffic dataset is used for forecasting by Artificial Neural Networks (ANN), which is a commonly used method in this research area. At first, statistical analysis is conducted to reveal the structure of the data such as seasonality, trend, etc. Then for the organized data, backpropagation ANN model is set up for forecasting the traffic flow. Finally, the forecast values are compared with the real data and forecasts using seasonal Autoregressive Integrated Moving Average (SARIMA) method. With the proposed ANN model, successful forecasts can be obtained.
Keywords: Traffic, forecast, artificial neural networks, SARIMA.

1. INTRODUCTION

Short-term traffic flow forecasting is one of the most known topics in the intelligent transportation systems. Traffic flow is usually forecasted by using historical and real-time data. With the emergence of Internet of Things (IoT) and therefore the rapid increase of data sources, realistic and effective solutions can be presented for short-term traffic problems. Real-time traffic flow data is collected from a wide range of data sources such as GPS, street cameras, social networks and online maps.

Traffic congestion, which is the most known traffic problem and causes material and immaterial damages, has a substantial effect on citizens living in the cities. If the traffic congestion can be forecasted before it happens, it can be easy to take precautions in short term. Therefore, using the real-time traffic flow data and effective forecasting algorithms the damages of the traffic congestions can be reduced. In this case, the forecasting models that provide the most effective and accurate results can provide substantial benefits.

In this study, we propose ANN models for the traffic flow forecasting depends on historical data. Seasonal Autoregressive Integrated Moving Average (SARIMA) method is also implemented to compare the performance of ANN implementation. Moreover, statistical analysis of the historical data is presented. The rest of this paper organized as follows. Some of the related literature about traffic flow forecasting will be reviewed in Section 2. ANN and SARIMA

* Corresponding Author: e-mail: zierzurum@eskisehir.edu.tr, tel: (222) 335 05 80

methods will be given briefly in Section 3 and 4 respectively. In Section 5, information about the datasets, statistical analysis of the data and the computational results will be explained. Finally, the results and the future works will be discussed in Section 6.

2. LITERATURE REVIEW

There are many different kinds of forecasting studies in traffic. In the literature traffic flow, volume, demand, speed, peak traffic volume, number of accidents, duration and severity of accidents are tried to be forecasted. Traffic volume, meteorological information, season, speed rate, vehicle type, age and gender of driver, information about road such as surface, slope and type, the type of accident, the cost of fuel and highway toll are some of the parameters which are used for the forecasts mentioned.

In this study, the literature about traffic flow forecasting in recent years is examined. To forecast the traffic flow, Abadi et al. [1] used an autoregressive model and least squares method using current and historical data. Annunziato et al. [2] proposed a hybrid modeling approach which combines Artificial Neural Networks and a simple statistical approach in order to provide a one hour forecast of urban traffic flow rates. In a similar manner, Chan et al. [3] developed a neural network based on an exponential smoothing method to enhance previously used neural network for traffic flow forecasting. Hosseini et al. [4] proposed a novel short-term traffic forecasting model using Multi-Layer Perceptron with Mutual Information and Feature Selection approach. Hou [5] focused on traffic flow forecasting in leisure farm areas using ANNs. In order to forecast real-time traffic flow state, Lu et al. [6] presented a simulated annealing genetic algorithm based fuzzy c-means algorithm. Lv et al. [7] presented a novel deep learning-based traffic flow method for traffic flow forecasting. Matas et al. [8] proposed a dynamic model to forecast traffic flow for tolled motorway in Spain. Moretti et al. [9] presented a hybrid model, which combines ANN, and a statistical approach in order to provide a one-hour forecast of urban traffic flow rates. Oh et al. [10] proposed an urban traffic flow forecasting system using a multifactor pattern recognition model, which combines Gaussian mixture model clustering with an ANN. For short-term traffic forecasting, Zhao et al. [11], proposed a long short-term memory (LSTM) network and also presented a novel algorithm which contains ARIMA algorithm [12]. Do et al. presented a survey about short-term traffic state prediction using network-based methods [13].

As seen from the literature, ANN models are proposed for traffic flow forecasting in general. In this study, we used two ANN models and seasonal ARIMA as a classical time series model to test the performances of these implementations on considered datasets. One of the models was developed with the classical backpropagation algorithm and the other model was implemented using Keras. The backpropagation algorithm was chosen because backpropagation is the most common technique used to train ANNs [13]. Also, Keras uses backpropagation when calculating the weights of ANN interconnections. In literature, there are relatively a few studies [14,15,16] about traffic forecasting in which Keras was used to implement neural networks. Since it is important to get fast and accurate forecasts in traffic, we thought that Keras could provide successful forecasts thanks to its features such as fast, easy-to-implement, modular and developable. Therefore, in this study, we aimed to show the strength of this implementation for studied traffic flow datasets.

3. ARTIFICIAL NEURAL NETWORKS

Inspired by biological systems, particularly by research into the human brain, Artificial Neural Networks (ANN) are able to learn from and generalize from experience. ANN provide an attractive alternative tool for both forecasting researchers and practitioners [17]. ANN is one of the commonly used method for forecasting and also used for pattern recognition, clustering, classification and optimization etc. Basically, a supervised learning approach is conducted in forecasting by ANN.

An ANN model basically consists of an input layer, a hidden layer and an output layer. Each layer is comprised of neurons that process the input signals and produce an output, while connections between the layers have a weight factor. ANN easily adjusts to any set of input-output patterns and through a robust training process forms a model function with the minimum possible error [18]. A basic ANN model is given in Figure 1.

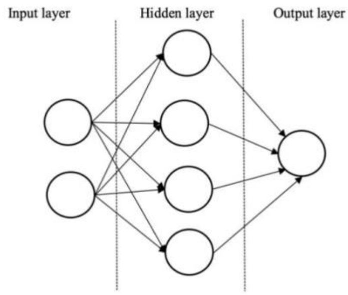

Figure 1. A basic ANN model

Backpropagation algorithm is the most famous algorithm, which is used for training of the feed-forward ANNs. The backpropagation algorithm looks for the minimum of the error function in weight space using the method of gradient descent. The combination of weights, which minimizes the error function, is considered to be a solution of the learning problem [19].

4. AUTOREGRESSIVE INTEGRATED MOVING AVERAGE METHOD

ARIMA is a famous forecast approach, which first introduced by Box and Jenkins [18]. In ARIMA models a non-stationary time series is made stationary by applying finite differencing of the data points [20]. As the parameters of ARIMA, p, d and q are integers greater than or equal to zero and refer to the order of the autoregressive, integrated, and moving average parts of the model respectively.

The ARIMA model turns into to seasonal ARIMA (SARIMA) model when there exist seasonal factors. A seasonal autoregressive notation (P) and a seasonal moving average notation (Q) will form the multiplicative process of SARIMA as $(p,d,q)(P,D,Q)_s$ where *s* shows the length of seasonal period [21]. A time series $\{Z_t \mid t = 1, 2, \dots, k\}$ is generated by SARIMA $(p,d,q)(P,D,Q)_s$ is given in Equation (1) [22]:

$$\phi_p(B)\Phi_p(B^s)(1-B)^d(1-B^s)^D Z_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t \tag{1}$$

where *p, d, q, P, D, Q* are integers, *s* is the season length;

$$\begin{gathered}\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p, \\ \Phi_P(B^s) = 1 - \Phi_s B^s - \Phi_{2s} B^{2s} - \cdots - \Phi_{Ps} B^{Ps}, \\ \theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_p B^p \text{ and} \\ \Theta_Q(B) = 1 - \Theta_s B^s - \Theta_{2s} B^{2s} - \cdots - \Theta_{Qs} B^{Qs}\end{gathered}$$

are polynomials in *B* of degree *p, q, P*, and *Q*. *B* is the backward shift operator, and ε_t is the estimated residual at time *t*. *d* is the number of regular differences, *D* is the number of seasonal differences; Z_t denotes the observed value at time *t*, $t = 1, 2, \dots, k$.

5. COMPUTATIONAL EXPERIMENTS

In this study, Eskisehir, Turkey and London are selected as testbeds. The traffic flow datasets of London Highways [23] include the data collected every 15 minutes from GPSs on the links between specific roads. Similarly, the traffic flow datasets of Eskisehir were collected every 15 minutes from specific intersections using fish-eye cameras. Both of the datasets include the flow counts of different types of vehicles individually (car, bus, minibus, pickup/panelvan/truck, tramway, bicycle, motorcycle etc). The 30 days of London traffic flow data and the 4 days of Eskisehir traffic flow data can be accessed and total flow values for each 15 minutes period are calculated.

To determine the variation of the time series, coefficient of variation is also calculated for each dataset. A coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean. The coefficient of variation represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from one another [24]. The size and coefficient of variation of each dataset are given in Table 1 below. As seen from Table 1, the variation of London datasets is quite high.

Table 1. The size and CV of datasets

Dataset	Size	Mean Flow Quantity	CV
London1	2976	230	0.669
London2	2881	498	0.664
London3	2977	349	0.753
London4	2880	362	0.868
London5	2976	273	0.814
Eskisehir1	56	93	0.242
Eskisehir2	72	88	0.160
Eskisehir3	72	44	0.207

In addition, regression analysis was conducted to examine the trend in London datasets. The trendlines are determined and evaluated statistically for each dataset. Figure 2 and the results of regression analysis clearly show that there is no trend in London1 and London2 and slight trendlines are detected in London3, London4 and London5.

Traffic flow is usually periodic, and the cycle is usually one day. Mostly, the peak hour of traffic is usually in the morning and late afternoon. The peak hour of traffic is usually in the early morning and late afternoon. The traffic flow variation trend of different days is usually similar. As a result, the seasonal property is a factor that can be considered [15]. The seasonal effect can be also observed in the datasets shown in Figure 2 prominently.

In order to evaluate the performance of proposed ANN method, a SARIMA model is implemented. Since the traffic flow has a seasonal property, a seasonal method is found more suitable for traffic flow forecasting.

ANN and SARIMA is implemented using Python version 3.5 with Statsmodels and Keras libraries. ANN is implemented using Keras API with TensorFlow backend. Keras is a high-level neural networks API, written in Python and capable of running on top of TensorFlow, CNTK, or Theano [25]. Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration [26].

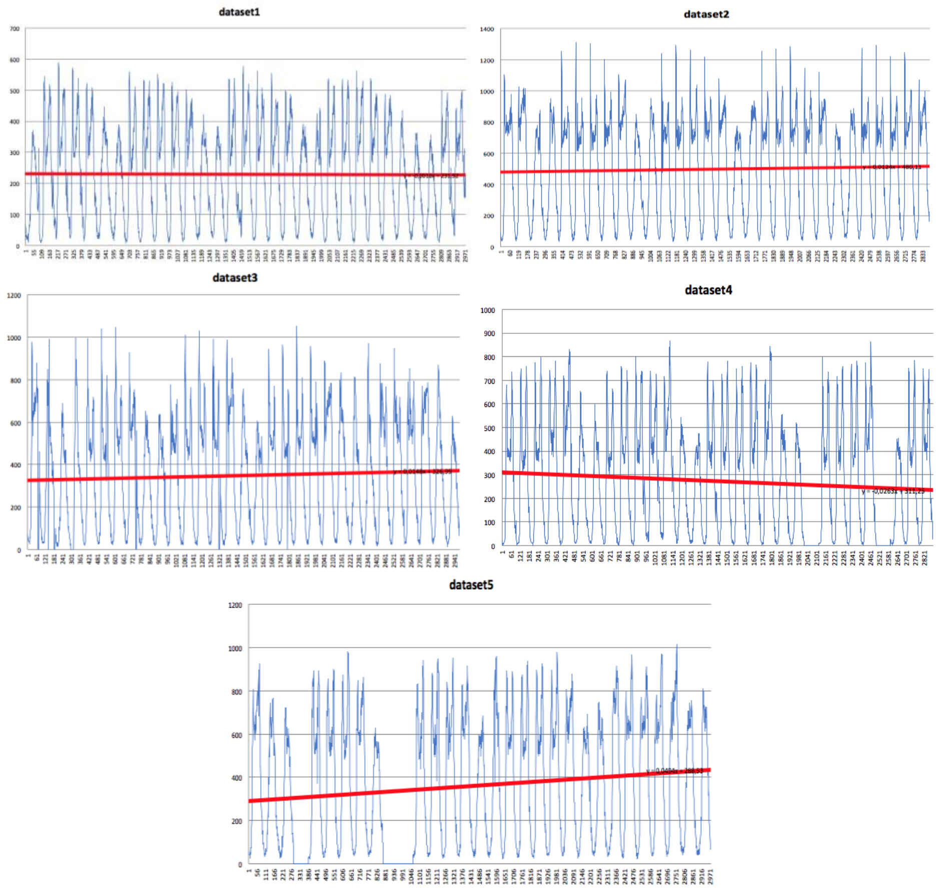

Figure 2. The trendlines of London datasets

Both of the ANN models have the same architectures. The implemented models have 3 layers: an input, a hidden and an output layers. The number of neurons in the hidden layer is tried as 6,7 and 8 neurons to find a more accurate model using validation dataset.

To tune the parameters of SARIMA, a grid search process is applied on *(p,d,q)* parameters. While the value of parameter *p* is changed between (0, 1, 2, 4, 6, 8, 10), the parameters *d* and *q* are changed in the range of [0,3] for Eskisehir datasets, the parameters *p*, *d* and *q* parameters are changed between 0 and 1 for London datasets because of the long running times. The seasonality parameters (*P*,*D*,*Q*) are fixed as (1,0,0) and the *s* parameter, length of seasonal period is changed according to dataset.

For both of the ANN models, 60% of each dataset is divided into training dataset, while 20% as validation and 20% as test. To validate and test the performance of the forecasting models, mean absolute error (MAE) and mean squared error (MSE) values calculated as using the formula in (2) and (3) respectively where *D* refers to actual value, *F* forecasted value and n number of forecasts:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|F_i - D_i| \quad (2)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(F_i - D_i)^2 \tag{3}$$

After training phase, validation and test phase are conducted for both of the ANN models. The MSE and MAE values for 6, 7 and 8 hidden neurons as a result of validation phase is given in Table 2.

Table 2. MAE and MSE values of validation phase for the ANN model using Keras

	n=6		n=7		n=8	
	MAE	MSE	MAE	MSE	MAE	MSE
London1	9.501	124.659	9.511	124.685	9.519	124.702
London2	15.961	391.451	15.977	391.878	15.925	390.621
London3	8.373	111.681	8.374	111.693	8.375	111.702
London4	3.794	19.329	3.807	19.455	3.859	20.460
London5	3.397	19.288	3.408	19.409	3.417	19.511
Eskisehir1	4.852	24.027	4.853	24.032	4.851	24.016
Eskisehir2	1.729	3.271	1.732	3.279	1.730	3.277
Eskisehir3	1.568	2.721	1.566	2.717	1.564	2.715

At the end of validation phase, the number of hidden neurons is determined specific to datasets.

The best test performance values of SARIMA and two ANN models using different parameter sets are given respectively in Table 3.

Table 3. MAE and MSE values of obtained forecasts using ARIMA and ANN models

	SARIMA		ANN (Keras)		ANN (Backpropagation)	
	MAE	MSE	MAE	MSE	MAE	MSE
London1	61.043	8229.428	9.877	135.146	76.537	10129.27
London2	65.261	8913.781	15.855	388.271	102.515	18808.136
London3	32.665	2101.481	8.791	120.468	77.504	12150.098
London4	48.472	4660.962	3.937	20.765	61.355	6137.646
London5	44.757	4783.681	3.327	19.007	54.187	6641.898
Eskisehir1	12.466	206.07	3.682	14.261	19.271	523.295
Eskisehir2	8.192	96.958	1.927	3.937	10.346	138.961
Eskisehir3	5.432	44.12	1.486	2.404	5.689	50.823

As seen from Table 2, the ANN model using gives better forecasts than the SARIMA model. Figure 3 and 4 also can be an example for the actual and obtained forecasted values. In parallel with the performance values, it can be seen that ANN using Keras gives more accurate forecasts to actual values than SARIMA from the graphs.

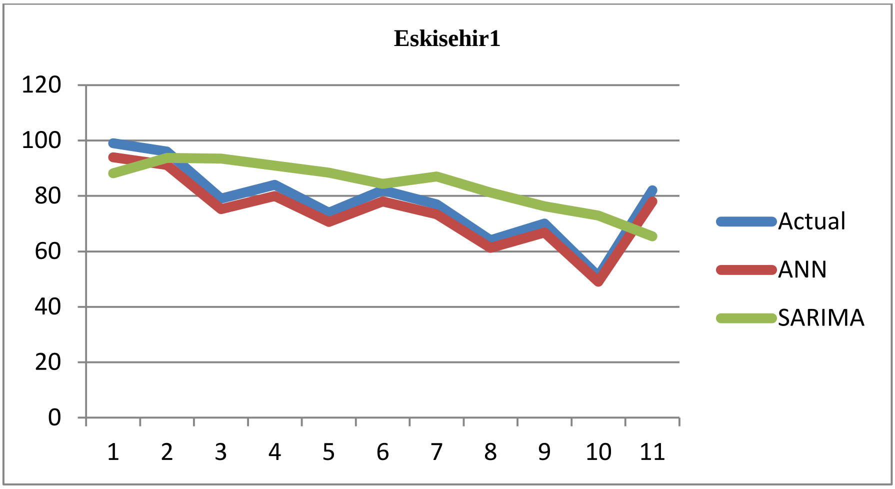

Figure 3. The comparison of actual and forecasted values for Eskisehir1 dataset

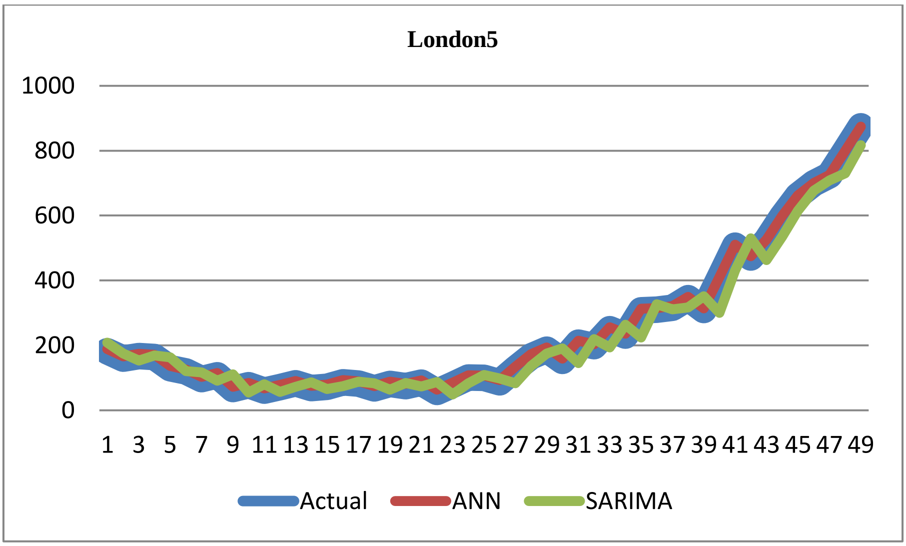

Figure 4. The comparison of actual and forecasted values for London5 dataset

Overtraining refers to the time of ANN training that may finally result in worse predictive ability of a network [27]. Therefore, the training process is carefully examined and iteration based graphics for training errors are created. Examples of these graphics are shown in Figure 5. As seen from Figure 5, training errors are decreasing in the training process and this confirms that there is no overtraining problem for ANN models.

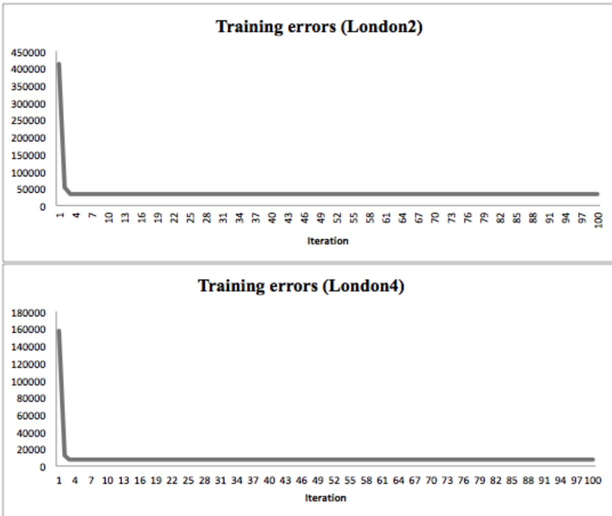

Figure 5. Training errors

6. CONCLUSIONS

The results of this forecasting study show that the ANN implementation with Pyhton Keras Library can come up with successful and fast results. To determine the efficiency of implemented model, it is needed to make more tests using larger datasets, which includes in streaming real-time data for intelligent transportation systems.

As seen from the literature review, the ANN models are enhanced with statistical techniques and evolutionary algorithms. From this point of view, novel hybrid models, which include ANN, can be developed to forecast traffic flow with streaming real-time data for future research. Since Keras is a deep learning library, a deep learning approach for traffic flow forecasting is also planned as future studies.

Besides these results, the SARIMA model has been found to be slow with large data, especially when the season length is long. Therefore, a novel SARIMA model can be generated for similar cases.

Finally, traffic flow is influenced by lots of parameters in real-life. Meteorological information, type of day, road and vehicle, traffic volume etc. should be taken into account in forecasting studies for traffic flow.

Acknowledgement

This study is supported by Anadolu University Scientific Research Projects Committee (AUBAP- 1709F506).

REFERENCES

[1] Abadi A., Rajabioun T., Ioannau P.A., (2015) Traffic Flow Prediction for Road Transportation Networks with Limited Traffic Data, *IEEE Transactions on Intelligent Transportation Systems*, 16, 653-662.

[2] Annunziato M., Moretti F., Pizzuti S., (2012) Urban Traffic Flow Forecasting Using Neural-Statistic Hybrid Modeling, *Soft Computing Models in Industrial and Environmental Applications*, 183-190.

[3] Chan K.Y., Singh J., Dillon T.S., Chang E., (2011) Traffic Flow Forecasting Neural Networks Based on Exponential Smoothing Method, *6th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 21-23 June 2011, Beijing, China.

[4] Hosseini S.H., Moshiri B., Rahimi-Kian A., Araabi B.N. (2012) Short-Term Traffic Flow Forecasting by Mutual Information and Artificial Neural Networks, *2012 IEEE International Conference on Industrial Technology, ICIT*, 19-21 March 2012, Athens, Greece.

[5] Hou C.I., (2013) Traffic Flow Forecasting in Leisure Farm Areas Using Artificial Neural Networks, *Przegląd Elektrotechniczny*, 89, 205-207.

[6] Lu H, Sun Z., Qu W., (2015) Big Data-Driven Based Real-Time Traffic Flow State Identification and Prediction, *Discrete Dynamics in Nature and Society*.

[7] Lv Y., Duan Y., Kang W., Li Z., Wang F.Y., (2015) Traffic Flow Prediction with Big Data: A Deep Learning Approach, *IEEE Transactions on Intelligent Transportation Systems*, 16, 865-873.

[8] Matas A., Raymond J.L., Ruiz A., (2012) Traffic forecasts under uncertainty and capacity constraints, *Transportation*, 39, 1-17.

[9] Moretti F., Pizzuti S., Panzieri S., Annunziato M., (2015) Urban traffic flow forecasting through statistical and neural network bagging ensemble hybrid modeling, *Neurocomputing*, 167, 3-7.

[10] Oh S., Kim Y., Hong J., (2015) Urban Traffic Flow Prediction System Using a Multifactor Pattern Recognition Model, *IEEE Transactions on Intelligent Transportation Systems*, 16, 2744-2755.

[11] Zhao Z., Chen W., Wu X., Chen P.C., Liu J., (2017) LSTM network: a deep learning approach for short-term traffic forecast, *IET Intelligent Transport Systems*, 11, 67-75.

[12] Zhao Z., Chen W., Yue H., Liu Z., (2016) A Novel Short-Term Traffic Forecast Model Based on Travel Distance Estimation and ARIMA, 2016 Chinese Control and Decision Conference, CCDC, 28-30 May 2016, Yinchuan, China.

[13] Do. L.N.N., Taherifar N., Vu H.L., (2018), Survey of neural network-based models for short-term traffic state prediction, *WIREs Data Mining Knowl Discov*.

[14] Azzouni A., Pujolle G., (2017), A long short-term memory recurrent neural network framework for network traffic matrix prediction. *arXiv preprint* arXiv:1705.05690.

[15] Yu R., Li Y., Shahabi C., Demiryurek U., Li Y., (2017), Deep Learning: A Generic Approach for Extreme Condition Traffic Forecasting, *Proceedings of, the 2017 SIAM International Conference on Data Mining*, 777-785.

[16] Wu Y., Tan H., Qin L., Ran B., Jiang Z., (2018), A hybrid deep learning based traffic flow prediction method and its understanding, *Transportation Research Part C: Emerging Technologies*, 90, 166-180.

[17] Zhang G., Patuwa B. E., Hu M.Y., (1998) Forecasting with artificial neural networks: The state of the art, *International Journal of Forecasting*, 14, 35-62.

[18] Kardakos E.G., Alexiadis M.C., Vagropoulos S.I., Simoglou C.K., Biskas P.N., Bakirtzis A.G., (2013) Application of Time Series and Artificial Neural Network Models in Short-term Forecasting of PV Power Generation, *48th International Universities' Power Engineering Conference*, *UPEC*, 2-5 September 2013, Dublin, Ireland.

[19] Rojas R., (1996), Neural Networks A Systematic Introduction, *Springer*.

[20] Adhikari R., Agrawal R.K., (2013) An Introductory Study on Time Series Modeling and Forecasting, *Lambert Academic Publishing.*

[21] Permanasari A.E., Hidayah I., Bustoni I.A., (2013) SARIMA (Seasonal ARIMA) Implementation on Time Series Forecast The Number of Malaria Incidence, *2013*

International Conference on Information Technology and Electrical Engineering, ICITEE, 07-08 October 2013, Yogyakarta, Indonesia

[22] Chen K.Y., Wang C.H., (2007) A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan, *Expert System with Applications*, 32, 254-264.

[23] Highways England, September, (2017) Online. Available: http://tris.highwaysengland.co.uk/detail/trafficflowdata.

[24] Coefficient of Variation – CV, (2017) Online. Available: http://www.investopedia.com/terms/c/coefficientofvariation.asp.

[25] Keras Documentation, (2017). [Online]. Available: https://keras.io/.

[26] Statsmodels's Documentation (2018) Online Avaliable: https://www.statsmodels.org/stable/index.html

[27] Tetko I.V., Livingstone D.J., Luik A.I., (1995), Neural Network Studies. 1. Comparision of Overfitting and Overtraining, *J. Chem. Inf. Comput. Sci*., 35, 826-833.