Australasian Transport Research Forum 2021 Proceedings
8-10 December, Brisbane, Australia
Publication website: http://www.atrf.info

Traffic Forecasting in a Freeway Corridor using Seasonal ARIMA Model

Mahmuda Akhtar[1*], Sara Moridpour[1], Majidreza Nazem[1]

[1] Civil and Infrastructure Engineering, School of Engineering, RMIT University, Australia

Email for correspondence: s3799862@student.rmit.edu.au

Abstract

Traffic congestion is becoming a critical problem in everyday life. The crucial need for a sustainable traffic forecasting method is becoming acute with time. This study develops an Auto Regressive Integrated Moving Average (ARIMA) model to estimate the short-term traffic forecasting in uninterrupted traffic flow using stationary sensor data. The transportation authorities can use the developed model to predict and avoid any traffic congestion incident by planning. In this paper, a major freeway section of Melbourne, Australia, is used as a case study. The model uses traffic volume data of 15-minutes intervals of 63 workdays from the Eastern freeway westbound corridor of Melbourne, Australia. Among 63 days, almost 50 days of data were used as the model's input variable, and the rest 13 days data was used to validate the developed model. The model's Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) values were as low as 0.28, 0.21, and 8.38%, respectively. The model was validated against the seasonal naïve model and found more effective than seasonal naïve.

Keywords: Traffic congestion; Forecasting; Seasonal ARIMA; Freeway.

1. Introduction

The importance of traffic forecasting is increasing every year, as the traffic congestion severity and its associated cost are rising worldwide. In Australia, traffic congestion costs in 2010 and 2015 were $12.8bn and $16.5bn, respectively, and will increase up to $37.3bn in 2030 if no planning is taken (BITRE 2015). Another study showed that Australia's major cities, e.g. Sydney and Brisbane, showed a traffic speed decrease of 3.6%, and Melbourne showed a decline of 8.2% from 2013 to 2018 (AAA 2018). Therefore, the urgency in forecasting traffic to reduce its associated cost and to increase road users' comfort is undeniable.

Availability of traffic data and the development of Artificial Intelligence (AI) methods have created new opportunities to predict traffic patterns. This opportunity enables transport authorities to make sustainable decisions to improve the traffic experience for dwellers' by reducing travel time and avoiding traffic incidents and traffic congestion.

This study aims to develop a seasonal Auto Regressive Integrated Moving Average (ARIMA) machine learning model to predict the traffic volumes for 15- minutes intervals into the future of a freeway corridor of Melbourne. The model is developed using historical data collected from a set of stationary sensors installed on the freeway. This data will be used to develop and verify the effectiveness of the machine learning model. The outcome of this study will assist

the transport authorities in predicting traffic congestion and establishing traffic congestion reduction strategies.

This paper summarises the literature providing the previous works in section 2. Section 3 describes the case study and collected data of the study. Section 4 summarises the methodology of the developed seasonal ARIMA model. Section 5 provides the results and discussions of the findings. Finally, the conclusions of the study, along with the future direction, is presented in Section 6.

2. Literature review

Many AI models have been used to forecast traffic using traffic data collected from different sources. In general, there are two traffic data sources, including stationary data and probe-vehicle data.

In recent years, traffic prediction has led to a growing research area, especially Artificial Intelligence (AI). With the introduction of big data by stationary sensors or probe data and the development of new models in AI in the last few decades, this research area has expanded extensively. Both short-term and long-term traffic prediction is made by evaluating different traffic parameters.

Traffic datasets used in different studies can be mainly divided into two classes, including stationary and probe data. Stationary data can be further divided into sensor data and fixed cameras (Cao and Wang, 2019, Kong et al., 2016, Yang et al., 2015, Zhang et al., 2019). On the other hand, probe data used in those studies were based on GPS data mounted on vehicles (Kong et al., 2016, Yang et al., 2015, Wang et al., 2015).

Akhtar and Moridpour (2021) have divided the AI models into three major categories. The categories include - Probabilistic Reasoning models, e.g., fuzzy logic (Onieva et al., 2012, Zhang et al., 2014, Lopez-Garcia et al., 2016), Hidden Markov Model (Zhao, 2015, Zheng et al., 2018, Zaki et al., 2019), Gaussian distribution (Yang, 2013, Zhu et al., 2019), Bayesian Network (Liu et al., 2014, Asencio-Cortés et al., 2016), Shallow Machine Learning (SML) algorithms, e.g., Artificial Neural Network (ANN) (Xu et al., 2019, Nadeem and Fowdur, 2018, Ito and Kaneyasu, 2017), regression models (Jiwan et al., 2015, Zhang and Qian, 2018, Jain et al., 2017, Alghamdi et al., 2019), decision tree (Wang et al., 2015, Liu and Wu, 2017), Support Vector Machine (SVM) (Tseng et al., 2018), and Deep Learning (DL) algorithms, e.g., Convolutional Neural Network (CNN) (Ma et al., 2017), Long short-term memory (LSTM) (Zhao et al., 2019). According to their study, the ARIMA model falls into the category of Shallow Machine Learning.

ARIMA is a popular model in the research area of forecasting. It has been applied in traffic forecasting for both the long and short term. The nature of the dataset plays an important role in model development. For example, Irhami and Farizal (2021) applied the ARIMA model for long term traffic number forecasting. As they used a long-term dataset, there was no noticeable seasonality. However, short-term traffic prediction using the whole day data can capture the seasonality and trend very well, making the decision-making process easier for the transport authority. Therefore, analysing seasonality is important. Alghamdi et al. (2019) used ARIMA for traffic congestion forecasting, applying 3- months data for the 1-hour interval. However, the authors did not consider seasonality in their study. Giraka and Selvaraj (2020) used only a 3-days dataset to predict the traffic volume of the intersection for peak- hours. Kumar and Vanajakshi (2015) also considered seasonality with 3-days worth of dataset. This study showed that the model

error decreased with the increment in the dataset. Therefore, it is important to model with a reasonably sized dataset to get better prediction results.

3. Case study

In this study, stationary sensor data was collected from a point of the Eastern freeway westbound corridor of Melbourne. It is the most crucial location on the freeway, which connects the Melbourne CBD entrance streets to the freeway. Therefore, this study deals with the temporal information of the location of interest on the freeway. Data was provided by the Department of Transport (Victoria) from September-November 2017. Data was of every minute traffic volume data for 24 hours. Figure 1 shows the study location of this research. A total of 131,040 raw data was collected for this study. This raw data was pre-processed and converted into 15-minutes interval data for the model development. The data processing is discussed in the next section.

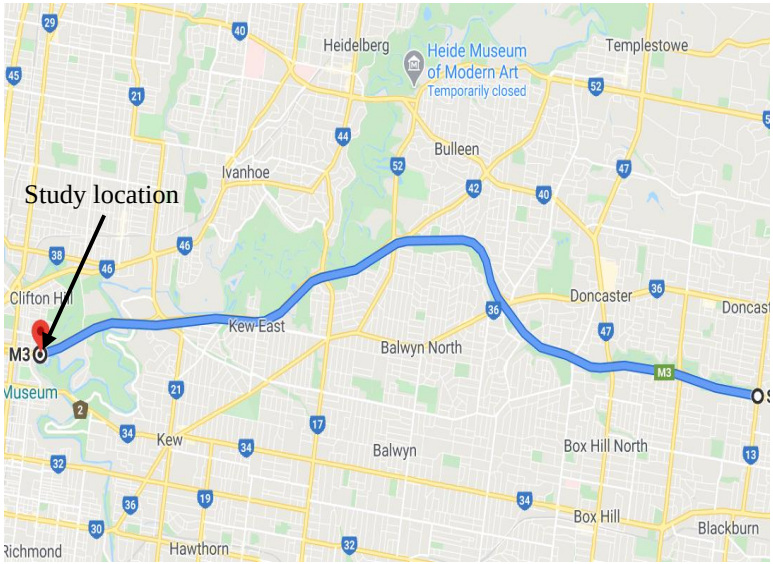

Figure 1: Location of the study area.

4. Methodology

The methodology of this research is presented in the following two subsections. The first section, Section 4.1, deals with the data preparation process. In the second section, Section 4.2, a detailed description and the methodology of the seasonal ARIMA model will be presented.

4.1. Data pre-processing

The dataset contained every minute's traffic data worth a total of 91 days. Among these 91 days, there were 26 weekends and two public holidays. We have excluded these 28 days from our dataset. After that, we pre-process the data by taking every 15-minutes of data. Therefore, there were 96 input variables (24 hr × 4 points/hr) each day. Finally, there were 6,048 traffic volume observations, with 69 missing data. This study ignores the missing data while preparing the final dataset. For model development, the data was split into 80% training set (4,838 data) and 20% testing and validation sets (1,210 data).

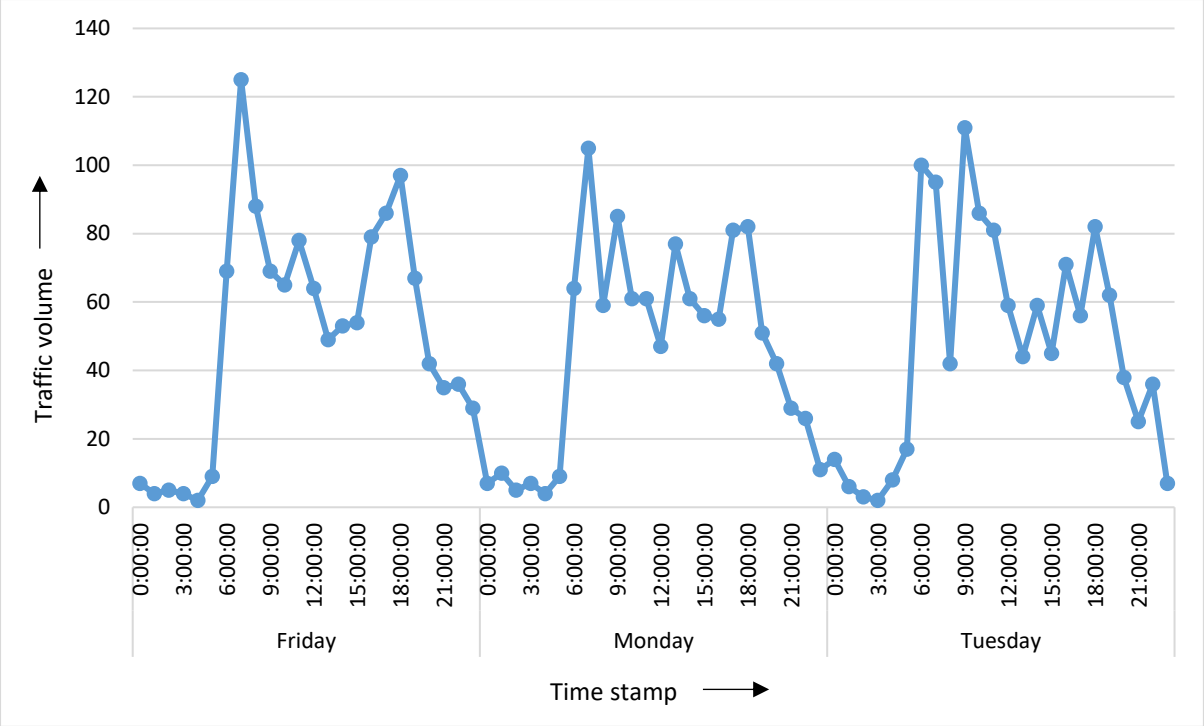

Figure 2: A sample plot of traffic volume data of three days.

Figure 2 shows a sample plot of three days of traffic volume data. A manual visualisation is done to find the trend or seasonality of the dataset. Figure 2 shows that there is no long term upward/downward trend; however, it shows seasonality every 24 hours. An important step in dataset preparation is to normalize the data. Our dataset showed a positive skewness. Therefore, we applied log-transformation to normalize the dataset.

4.2. Development of seasonal ARIMA

Seasonal ARIMA model development for short-time traffic volume prediction consists of three main steps. The first step is to plot the time-series data to identify the order of the parameters. The second step is to estimate the model parameters by diagnostic checking. The final step is to forecast the future traffic volume and model validation.

4.2.1. Parameter order identification

After converting the dataset into a time series, the first step to develop an ARIMA model is to make the stationary dataset. Making the stationary dataset means there cannot be any trend or seasonality in the dataset. A non-seasonal ARIMA model consists of three parameters- p, d, and q. p is the Auto Regressive (AR) order, q is the Moving Average (MA) order, and d refers to the non-seasonal difference. If there is a visible long-term trend, a non-seasonal differencing can make the dataset stationary. However, for the dataset with visible seasonality, both seasonal and non-seasonal parameters need to be incorporated. This modelling is called seasonal ARIMA modelling. A seasonal ARIMA can be presented as ARIMA (p, d, q) × $(P, D, Q)_S$, where P, D, Q are seasonal AR, differencing, and MA parameters and S is the time lag for seasonality. Both seasonal and non-seasonal differencing are needed for the time-series data with visual trends and seasonality. However, to avoid the unnecessary level of dependency, a maximum of two differencing is advisable (Giraka and Selvaraj, 2020).

As seen from Figure 2, there is no visible trend, but seasonality is present. Only seasonal differencing at lag 96 (y_t-y_{t-96}) was necessary to make the dataset stationary. To check whether

the dataset became stationary or not, four tests, including, Augmented Dickey-Fuller Test, Phillips-Perron Unit Root Test, KPSS Test, and Box-Ljung test, were done. These tests confirmed the stationarity of the dataset. Table 1 shows the values of these tests.

Table 1: Results from Unit Ratio Tests.

Unit Ratio Tests	Estimated P values	Required P values
ADF test	0.01	0.01
PP test	0.01	0.01
KPSS test	0.1	0.1
Box-Ljung test	< 2.2e-16	< 0.05

Next, for the differenced time series, the Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) was plotted to find the order of AR and MA. In Figure 3, the ACF plot shows a gradual decrease towards zero, and there is a significant correlation at the seasonal lag of 96. It suggests an AR process in the non-seasonal part and a possible MA (1) process in the seasonal part of the model. Now, from the PACF plot, the order of AR can be found. There are six significant non-zero correlations at the early lag. Therefore, the possible orders of the model are ARIMA (6,0,0) × $(0,1,1)_{96}$, ARIMA (5,0,0) × $(0,1,1)_{96}$, ARIMA (4,0,0) × $(0,1,1)_{96}$, ARIMA (3,0,0) × $(0,1,1)_{96}$, ARIMA (2,0,0) × $(0,1,1)_{96}$, and ARIMA (1,0,0) × $(0,1,1)_{96}$.

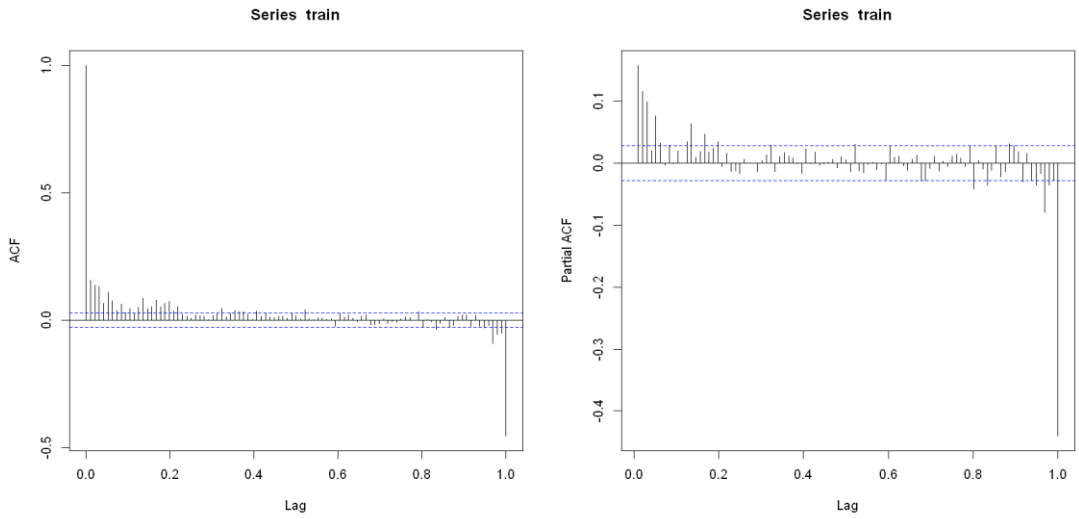

Figure 3: ACF and PACF plot of differenced series.

4.2.2. Model parameters estimation

Model development is done in Rstudio for the selected parameter orders for the training dataset. To estimate the best fitted model, Akaike's Information Criteria (AIC) was used. It is a prediction error estimator. Therefore, the smaller the value of AIC indicated the more accurate model. AIC uses the maximum likelihood method. AIC can be defined as follows:

$$AIC = 2k - 2\ln(\hat{L}) \quad (1)$$

Where, k is the estimated parameter numbers in the model and $\hat{L}$ represents the maximum likelihood function value for the model.

Table 2 shows the results of the models. It is seen that ARIMA $(6,0,0) \times (0,1,1)_{96}$ has the lowest AIC value among all the identified models. Therefore, ARIMA $(6,0,0) \times (0,1,1)_{96}$ is chosen as the best performing model.

Table 2: Seasonal ARIMA model parameters.

Model	Type	Parameters	Value	AIC
$(6,0,0)\times(0,1,1)_{96}$	Non-Seasonal AR	AR1	0.15	1021.93
		AR2	0.12	
		AR3	0.10	
		AR4	-0.01	
		AR5	0.07	
		AR6	0.05	
	Seasonal MA	MA1	-0.89	
$(5,0,0)\times(0,1,1)_{96}$	Non-Seasonal AR	AR1	0.15	1030.76
		AR2	0.12	
		AR3	0.11	
		AR4	0.00	
		AR5	0.07	
	Seasonal MA	MA1	-0.89	
$(4,0,0)\times(0,1,1)_{96}$	Non-Seasonal AR	AR1	0.15	1053.69
		AR2	0.13	
		AR3	0.11	
		AR4	0.01	
	Seasonal MA	MA1	-0.89	
$(3,0,0)\times(0,1,1)_{96}$	Non-Seasonal AR	AR1	0.16	1052.37
		AR2	0.13	
		AR3	0.12	
	Seasonal MA	MA1	-0.89	
$(2,0,0)\times(0,1,1)_{96}$	Non-Seasonal AR	AR1	0.17	1114.96
		AR2	0.15	
	Seasonal MA	MA1	-0.88	
$(1,0,0)\times(0,1,1)_{96}$	Non-Seasonal AR	AR1	0.16	1225.49
	Seasonal MA	MA1	-0.90	

4.2.3. Verification of the model

The selected model is then used to forecast the future of the time series. This predicted value was then compared with the observed testing data for verification. The forecasted value was also validated using three model performance indicators. The performance indicators are provided below:

Root Mean Square Error, $RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\bar{Y}_i - Y_i)^2}{n}}$ (2)

Mean Absolute Error, $MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \bar{Y}_i|$ (3)

Mean Absolute Percentage Error, $MAPE(\%) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \bar{Y}_i}{Y_i}\right| \times 100$ (4)

Here, $\bar{Y}$ represents the original value and Y is the predicted value, n is the number of instances. The forecasted result was also compared seasonal naïve model for validation.

5. Results and discussion

The selected model ARIMA (6,0,0) × $(0,1,1)_{96}$ was applied for testing the dataset length. Therefore, the model predicts 1,210 values of traffic volume. Figure 4 shows the observed and forecasted traffic data. It is seen that the forecasted values are not very different from the observed values.

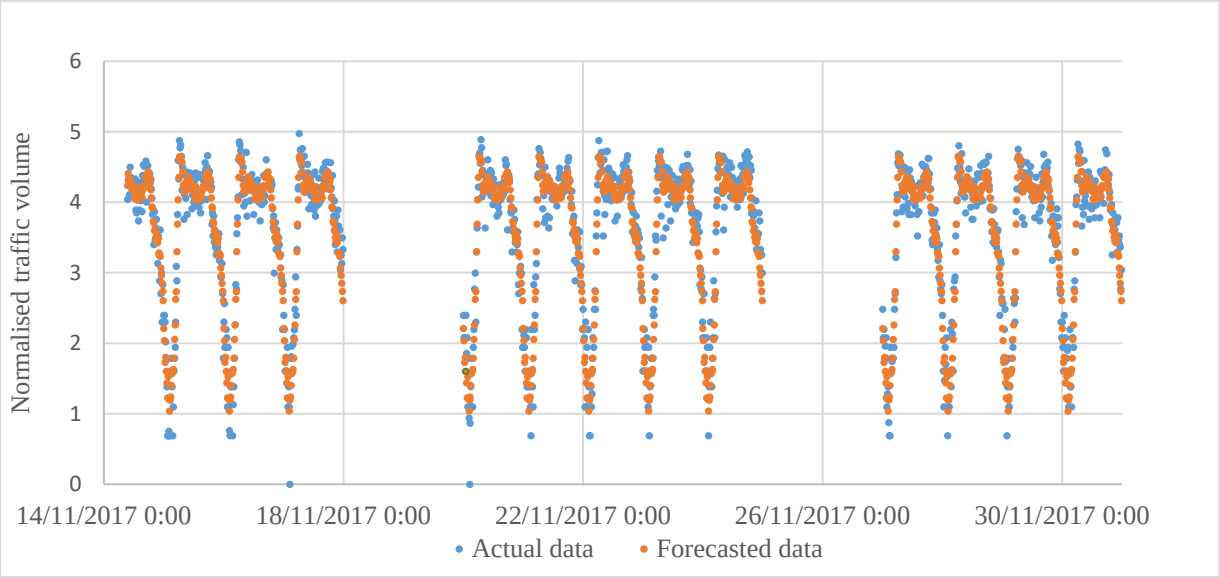

Figure 4: Observed versus predicted traffic volume data plot for seasonal ARIMA model.

To understand how effectively the model is working, a comparison with models from other studies were made. The results found from the seasonal ARIMA model are summarised in Table 3. The results obtained from the seasonal ARIMA model was compared with the seasonal naïve model for the same forecasting dataset. It is observed that the seasonal ARIMA model outperforms the seasonal naïve model in terms of performance indicators. Also, according to Lawrence & Klimberg (1982), if a forecasted result produces a MAPE value of less than 10%, the model can be taken as highly accurate. A MAPE value between 10% and 20% is considered good, MAPE of 21%-50% is reasonable, and MAPE of more than 51% is taken as inaccurate modelling. Therefore, according to this scale, the proposed model is highly accurate as a time series forecasting method.

Table 3: Performance indicator results of seasonal ARIMA and seasonal naive model.

Model	Performance indicator			Accuracy (Lawrence & Klimberg, 1982)
	RMSE	MAE	MAPE	
Seasonal ARIMA	**0.28**	**0.21**	**8.38%**	High

Seasonal Naïve	0.37	0.28	11.75%	Good

6. Conclusions and future research direction

This study developed an efficient short-term traffic volume prediction method using a seasonal ARIMA model. Traffic volume data of every minute were collected from the Department of Transport (Victoria) for three months. The collected data were pre-processed and converted into 15-minute interval data. From the visual inspection of the dataset, it was evident that there was a clear 24-hour seasonality with no long term upward/downward trend. Therefore, instead of ARIMA, the seasonal ARIMA method was adopted. After that, from a range of identified models, the most efficient model was chosen using the widely adopted maximum likelihood method. Finally, to validate the developed model, three performance indicators, namely- RMSE, MAE, and MAPE, were determined and compared with the seasonal naïve model. The proposed model shows a very promising result and outperforms seasonal naïve model results for every indicator.

The future direction of the study is to investigate whether including more parameters, e.g. traffic speed improves the predictability of the forecasting model. In future, newly developed state-of-the-art models will be applied to find a more accurate forecasting model.

7. Acknowledgment

The authors would like to thank the Department of Transport (Victoria) for providing this research with the traffic data.

8. Reference

Akhtar, M. & Moridpour, S. 2021. A Review of Traffic Congestion Prediction Using Artificial Intelligence. *Journal of advanced transportation*, 2021.

Alghamdi, T., Elgazzar, K., Bayoumi, M., Sharaf, T. & Shah, S. Forecasting traffic congestion using ARIMA modeling. 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), 2019. IEEE, 1227-1232.

Asencio-Cortés, G., Florido, E., Troncoso, A. & Martínez-Álvarez, F. 2016. A novel methodology to predict urban traffic congestion with ensemble learning. *A Fusion of Foundations, Methodologies and Applications*, 20**,** 4205-4216.

Cao, W. & Wang, J. Research on traffic flow congestion based on Mamdani fuzzy system. AIP Conference Proceedings, 2019.

Giraka, O. & Selvaraj, V. K. 2020. Short-term prediction of intersection turning volume using seasonal ARIMA model. *Transportation letters*, 12**,** 483-490.

Irhami, E. A. & Farizal, F. 2021. Forecasting the Number of Vehicles in Indonesia Using Auto Regressive Integrative Moving Average (ARIMA) Method. Bristol: IOP Publishing.

Ito, T. & Kaneyasu, R. 2017. Predicting traffic congestion using driver behavior. *International Conference on Knowledge Based and Intelligent Information and Engineering*

Systems, KES2017. Marseille, France: Procedia Computer Science.

Jain, S., Jain, S. S. & Jain, G. 2017. Traffic Congestion Modelling Based on Origin and Destination.

Jiwan, L., Bonghee, H., Kyungmin, L. & Yang-Ja, J. 2015. A Prediction Model of Traffic Congestion Using Weather Data.

Kong, X., Xu, Z., Shen, G., Wang, J., Yang, Q. & Zhang, B. 2016. Urban traffic congestion estimation and prediction based on floating car trajectory data. *Future Generation Computer Systems*, 61**,** 97-107.

Kumar, S. V. & Vanajakshi, L. 2015. Short-term traffic flow prediction using seasonal ARIMA model with limited input data. *European transport research review*, 7**,** 1-9.

Liu, Y., Feng, X., Wang, Q., Zhang, H. & Wang, X. 2014. Prediction of Urban Road Congestion Using a Bayesian Network Approach. *Procedia - Social and Behavioral Sciences,* 138**,** 671-678.

Liu, Y. & Wu, H. Prediction of road traffic congestion based on random forest. 2017 10th International Symposium on Computational Intelligence and Design (ISCID), 2017. IEEE, 361-364.

Lopez-Garcia, P., Onieva, E., Osaba, E., Masegosa, A. D. & Perallos, A. 2016. A Hybrid Method for Short-Term Traffic Congestion Forecasting Using Genetic Algorithms and Cross Entropy. *IEEE Transactions on Intelligent Transportation Systems,* 17**,** 557-569.

Ma, X., Dai, Z., He, Z., Ma, J., Wang, Y. & Wang, Y. 2017. Learning Traffic as Images: A Deep Convolutional Neural Network for Large-Scale Transportation Network Speed Prediction. *Sensors (Basel, Switzerland),* 17.

Nadeem, K. M. & Fowdur, T. P. 2018. Performance analysis of a real-time adaptive prediction algorithm for traffic congestion. *Journal of Information and Communication Technology,* 17**,** 493-511.

Onieva, E., Milanes, V., Villagra, J., Perez, J. & Godoy, J. 2012. Genetic optimisation of a vehicle fuzzy decision system for intersections.

Tseng, F.-H., Hsueh, J.-H., Tseng, C.-W., Yang, Y.-T., Chao, H.-C. & Chou, L.-D. 2018. Congestion Prediction With Big Data for Real-Time Highway Traffic. *IEEE Access,* 6**,** 57311-57323.

Wang, J., Mao, Y., Li, J., Zhang, X. & Wen-Xu, W. 2015. Predictability of Road Traffic and Congestion in Urban Areas. *PLoS One,* 10**,** e0121825.

Xu, Y., Shixin, L., Keyan, G., Tingting, Q. & Xiaoya, C. 2019. Application of Data Science Technologies in Intelligent Prediction of Traffic Congestion. *Journal of Advanced Transportation,* 2019.

Yang, Q., Wang, J., Song, X., Kong, X., Xu, Z. & Zhang, B. Urban traffic congestion prediction using floating car trajectory data. International Conference on Algorithms and Architectures for Parallel Processing, 2015. Springer, 18-30.

Yang, S. 2013. On feature selection for traffic congestion prediction. *Transportation Research Part C,* 26**,** 160-169.

Zaki, J. F., Ali-Eldin, A., Hussein, S. E., Saraya, S. F. & Areed, F. F. 2019. Traffic congestion prediction based on Hidden Markov Models and contrast measure. *Ain Shams Engineering Journal*.

Zhang, P. & Qian, Z. 2018. User-centric interdependent urban systems: Using time-of-day electricity usage data to predict morning roadway congestion. *Transportation Research Part C,* 92**,** 392-411.

Zhang, W., Yu, Y., Qi, Y., Shu, F. & Wang, Y. 2019. Short-term traffic flow prediction based on spatio-temporal analysis and CNN deep learning. *Transportmetrica A: Transport Science,* 15**,** 1688-1711.

Zhang, X., Onieva, E., Perallos, A., Osaba, E. & Lee, V. C. 2014. Hierarchical fuzzy rule-based system optimised with genetic algorithms for short term traffic congestion prediction. *Transportation Research Part C: Emerging Technologies,* 43**,** 127-142.

Zhao, H., Jizhe, X., Fan, L., Zhen, L. & Qingquan, L. 2019. A Peak Traffic Congestion Prediction Method Based on Bus Driving Time. *Entropy,* 21**,** 709.

Zhao, J. 2015. Research on Prediction of Traffic Congestion State. *MATEC Web of Conferences.* Les Ulis: EDP Sciences.

Zheng, Y., Li, Y., Own, C.-M., Meng, Z. & Gao, M. 2018. Real-time predication and navigation on traffic congestion model with equilibrium Markov chain. *International Journal of Distributed Sensor Networks,* 14.

Zhu, L., Krishnan, R., Guo, F., Polak, J. & Sivakumar, A. Early identification of recurrent congestion in heterogeneous urban traffic. 2019 IEEE Intelligent Transportation Systems Conference (ITSC), 2019. IEEE, 4392-4397.