ESTEEM Academic Journal
Vol. 14, December 2018, 66-76

MODELLING ROAD ACCIDENTS IN MALAYSIA

Aida Syamilah Mohd Nasir[1], Haziq Hamizan Husainizam[1], Nurul Sabrina Halim[1], Isnewati Ab Malek[1] and Haslinda Ab Malek[1]
[1] *Faculty of Computer and Mathematical Sciences, UiTM Cawangan Negeri Sembilan, Kampus Seremban, Negeri Sembilan, Malaysia*
**corresponding author: [4]isnewati@ns.uitm.edu.my*

ARTICLE HISTORY

Received
1 October 2018

Accepted
5 December 2018

Available online
30 December 2018

ABSTRACT

Safety and accident issues are considered important problems in the world. Road accident issues would have more conspicuous countenance in Malaysia. Almost every day there will be news about road accident, whether on television, newspapers and internet. Therefore, the aim of this study is to understand the general trend of road accidents in Malaysia, to estimate the best parameters of the Box-Jenkins Model and then to find the best-fitted model for road accidents in Malaysia. A monthly accident data from 2007 to 2016 was obtained form the Malaysia Institude of Road Safety (MIROS). This study applies Autoregressive Integrated Moving Average (ARIMA) time series model to study the trend of road accidents in Malaysia. Based on the result obtained, the trend of the number of road accidents in Malaysia are increasing. The finding also showed that by comparing the AIC and BIC value, ARIMA(1,1,1) was identified as the best model. This study contributes to the implementation of road safety in order to reduce the increasing trend of road accidents in Malaysia.

Keywords: *road accidents; box-jenkin; ARIMA*

1. INTRODUCTION

Malaysia is one of the developing countries and the economy in this country is growing rapidly. Therefore, the number of road users have sharply risen in the last decade. There is no denying that road accidents in Malaysia have become a normal phenomenon as the number increased by 3.5% from 2008 to 2017. Road accidents are measured in terms of the number of persons injured and death due to road accidents. Accidents on the road can involve a range of scenarios. The causal factors for road accidents include reckless driving or ignoring traffic rules instead of the faulty road or vehicle condition.

Malaysia has been ranked as one of the top three nations in the world with the deadliest road. The World Health Organization (WHO) [8] indicates that Malaysia is among emerging countries with the riskiest roads after Thailand and South Africa. Road accident also is a major cause of death and injuries in Malaysia. In addition, Malaysian Institute of Road Safety Research (MIROS) has reported that the number of accidents and deaths have been on the rise from 1999 until 2015. Most of these road accidents have been attributed to over speeding and wrong overtaking by the drivers.

p-ISSN 1675-7939; e-ISSN 2289-4934

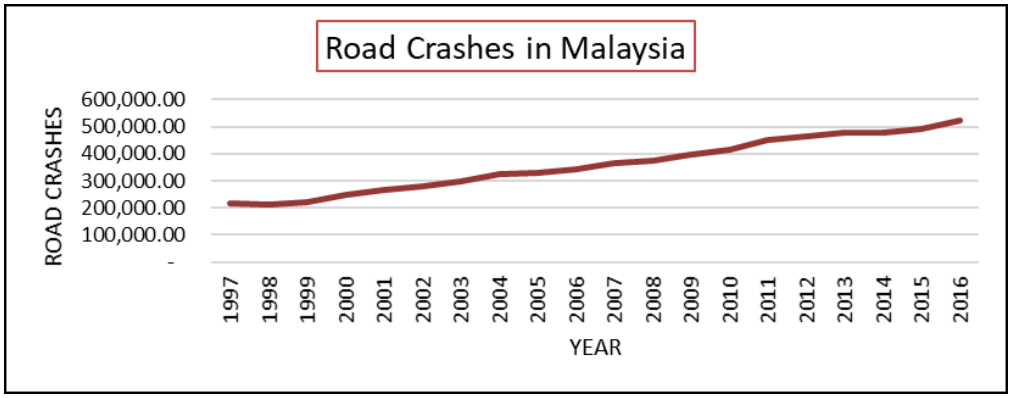


Figure 1: Number of Road Crashes in the Year 1997 until 2016 in Malaysia.

Figure 1 shows the number of road crashes from the year 1997 to 2016. The number of road crashes in Malaysia increased from 215,632 in 1997 to 521,466 in 2016. The accident prediction model will be expected to help other researchers in the studies to contribute to reducing the number of accidents in Malaysia. Therefore, the objectives of this study are to determine the trend of the road accidents in Malaysia, to estimate the best parameters of Box-Jenkins Model and Poisson distribution and to determine the best-fitted model of road accidents in Malaysia. It is hoped that this study can be beneficial to the road authorities in reducing the rate of accident cases in Malaysia.

1.1 Literature Review

This study involved the method of Autoregressive Integrated Moving Average (ARIMA). ARIMA is identified to be the suitable model for road accident data. In a study done by Avuglah[2] it showed that road accidents are increasing in Ghana and ARIMA (0, 2, 1) was identified as the best model to make a five-year forecast. According to Rohayu [7], ARIMA takes into account the previous observation and the past errors to observe patterns and make predictions. Based on their study, Poisson and Negative Binomial distribution seem to be inferior to the ARIMA model and the result showed that the best model for predicting Malaysian road fatalities was ARIMA (0, 1,1). This finding is supported by Kumar [5] which reported that the time series analysis based techniques like ARIMA are one of the most accurate methods for the prediction of traffic flow compared to historic average and naive model.

According to Balagun [3], the authors found that ARIMA (3,1,1) gave the lowest MSE value and MA (0,1,2) gave the lowest AIC value. Hence, ARIMA (3,1,1) and MA (0,1,2) were the best model for fitting and forecasting road accident data in Nigeria. The two criteria used to determine which of the models fits the best is the Akaike information criterion (AIC) and Mean Square Error (MSE).

Furthermore, Mahmoud [6] in his study used AIC and Bayesian Information Criteria (BIC) to determine the best fitted model. According to this the model with the least AIC value will be selected. He entertained nine tentative ARMA models and chose that model which has the minimum AIC. The result found that the lowest AIC and BIC values are for the ARIMA (1,0,0) (p=1, d=0 and q=0) and henceforth this model can be best model for making forecasts for future values.

p-ISSN 1675-7939; e-ISSN 2289-4934

2. METHODOLOGY

2.1 Source of Data

The sources of data for this study was secondary data. The monthly of road accident for the years 2007 through to 2016 was compiled by Malaysian Institute Road Safety (MIROS). In handling this study, E-views software packages have been used to analyze the road accident data in Malaysia. This software packages are useful to analyze a large volume of data.

2.2 The Trend Pattern of Road Accidents

Alias [1] in his book stated that the trend component in business or economic time series data describe the general upward or downward movements that characterized all economic and business activities which are usually found in dynamic economic and business environments. In short, the trend represents the long-run growth or decrease over time. The simplest method to identify the trend is to plot a straight line through the points on the graph.

2.2 Box-Jenkin Method

Box-Jenkins is a method involving the process of identifying, fitting and checking ARIMA models together with time series data. The mathematical formulation of the ARIMA has *(a, d, b)* form where *a*, *d*, and *b* are integers greater than or equal to zero and refer to the order of the autoregressive (*a*), the order of differencing (*d*) and the order of moving average (*b*). Thus, the equation can be written as in Equation (1).

$$\mu + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \ \ldots + \theta_b\varepsilon_{t-b} = x_t + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ \ldots + \phi_a x_{t-a} + \varepsilon \qquad (1)$$

According to Alias [1], the order of differencing is defined as the number of times need to be differenced to achieve stationary. Hence, if the difference between the current value, x_t, and the preceding value, x_{t-1} is taken and the result is stationary, then this is defined as the first order differencing. If the first difference is not achieved, then the second order differencing is performed.

2.2.1 The Stages in ARIMA Model Development

The basis of the Box-Jenkins modelling approach consists of three main stages. There are Model Identification, Model Estimation and Validation, and Model Application.

In Model Identification the step is to distinguish the class of the most suited model to be applied to the data set. After picking out models from ACF and PACF, run certain statistical procedures to take the best fitted model.

In Model Estimation and Validation, the portion is the work of selecting the best model for prediction purposes. The foremost target is the fitted values should be as near as possible to the real values. The second objective is that the models should involve the least possible parameters consistent with a good model fit.

p-ISSN 1675-7939; e-ISSN 2289-4934

In the end, the final phase in the development of ARIMA model is called as Model Application. At this stage, once all criteria are met and the model is corroborated to be fit, it is then ready to be used for prediction purposes.

2.2.2 Assumptions of Box-Jenkins

The application of Box Jenkins models assumes that the observation of the time series data is stationary. The data series is regarded as stationary if it fluctuates randomly around some fixed values, generally either around the mean value or constant value or even zero value. More specifically, if it does not show growth or decline over time, the series is said to be stationary.

2.2.3 Model Identification

The foremost step is to distinguish the class of model that is most suitable to be utilized to the given data set in the application of the Box-Jenkins methodology. This process is based on historical data by computing, analyzing and plotting various statistics. In order to select the best fitted model, one needs to run several models and by applying certain statistical test procedures.

Once the stationary condition has been attained, the next step is to test whether there exists any discernible form of the data series through the ACF and PACF. The order of the AR(a) equals the number of significant spikes in the PACF. Meanwhile, the order of the MA(b) equals the number of significant spikes in the ACF. The process to decide on the value of a and b in ARIMA model is difficult. Hence, it is worth considering several possible models in order to minimize the chance of not picking the most appropriate model form. To determine the best fitted model, one needs to use several statistical measures such as AIC/BIC or the Box-Pierce (Ljung-Box) statistic.

2.2.4 Performance Evaluation

The performance of the model is evaluated using few statistical measures such as Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC).

Akaike's Information Criteria (AIC) is to quantify the penalty on the likelihood for each additional term included in the model. In order to calculate AIC, the formula is as follows,

$$\text{AIC} = e^{\frac{2k}{T}\frac{\sum_{t=1}^{T} e_t^2}{T}} \qquad \text{where } k = a + b + A + B \tag{2}$$

a and b are the terms for AR and MA parts respectively, while A and B are the seasonality part of the ARIMA model with the existence of the seasonality factors. T is the total number of observations in the data series. In other words, a model is conceived as possessing a better fit among all other competing models if the value of its AIC is the smallest.

Bayesian Information Criteria (BIC) also known as Schwarz Criterion (SBC). The aim of this statistic measure is to select the model that achieve the most accurate out-of-sample forecasts by balancing between the model complexity and goodness of fit.

p-ISSN 1675-7939; e-ISSN 2289-4934

It can be calculated by using this formula;

$$\text{BIC} = T^{\frac{k}{T}} \frac{\sum_{t=1}^{T} e_t^2}{T} \tag{3}$$

where k = the number of parameters in estimated model
T = the number of observations in the series.

Likewise, to the AIC, the best example is the one that delivers the smallest BIC value.

2.2.5 Model Application

The main objective of the Box-Jenkins model is to provide a tool that can be used to generate and estimate the forecast value of a particular time series data. The generated value can be a single value. Repetitive model formulation and estimation is necessary to find the best final model. The process of formulating and estimating the models should be followed by regular re-evaluation and refine of the model estimated until the final model form. The final or best model is chosen based on the model's forecasting performance (smallest value of AIC and BIC).

3. RESULTS AND FINDINGS

3.1 The Trend Pattern of Road Accidents in Malaysia

The foremost aim of this study is to trace the trend pattern of the road accidents in Malaysia from 2007 until 2016. The data around the road accident shown in Figure 2 indicated that the number of road crashes in Malaysia increased from 215,632 in 1997 to 521,466 in 2016. In short, the trend pattern of the road accidents in Malaysia represents the long-run growth over the year. Referring to Figure 2, the equation showed that every month, the average number of road accidents in Malaysia increased by 120 cases.

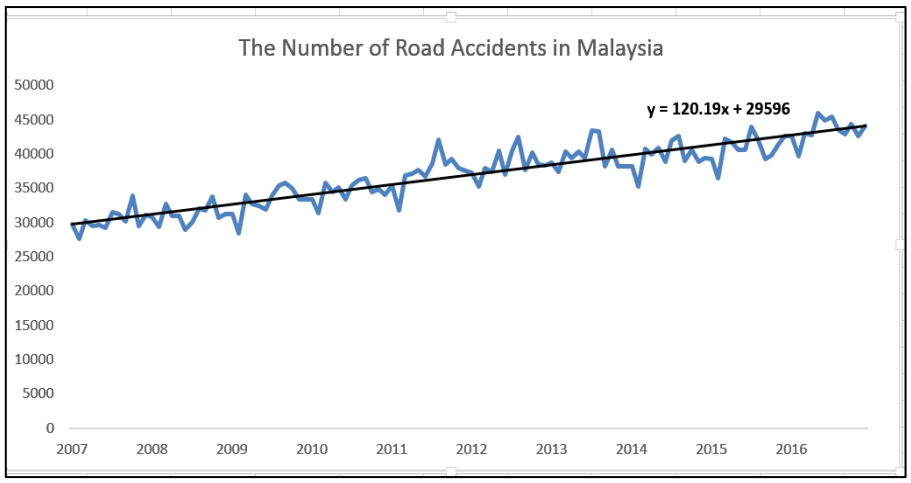
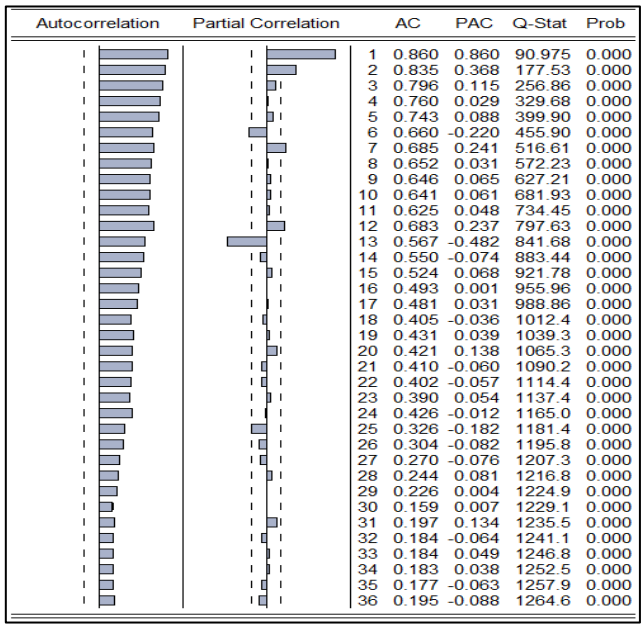


Figure 2: The Trend Pattern of the Road Accidents in Malaysia

p-ISSN 1675-7939; e-ISSN 2289-4934

3.2 Analysis of Box-Jenkins Model

Figure 3 below clearly shows that the correlogram of original data of ACF value decreases slowly and PACF value shows that there is a large significant spike at lag one.



	AC	PAC	Q-Stat	Prob
1	0.860	0.860	90.975	0.000
2	0.835	0.368	177.53	0.000
3	0.796	0.115	256.86	0.000
4	0.760	0.029	329.68	0.000
5	0.743	0.088	399.90	0.000
6	0.660	-0.220	455.90	0.000
7	0.685	0.241	516.61	0.000
8	0.652	0.031	572.23	0.000
9	0.646	0.065	627.21	0.000
10	0.641	0.061	681.93	0.000
11	0.625	0.048	734.45	0.000
12	0.683	0.237	797.63	0.000
13	0.567	-0.482	841.68	0.000
14	0.550	-0.074	883.44	0.000
15	0.524	0.068	921.78	0.000
16	0.493	0.001	955.96	0.000
17	0.481	0.031	988.86	0.000
18	0.405	-0.036	1012.4	0.000
19	0.431	0.039	1039.3	0.000
20	0.421	0.138	1065.3	0.000
21	0.410	-0.060	1090.2	0.000
22	0.402	-0.057	1114.4	0.000
23	0.390	0.054	1137.4	0.000
24	0.426	-0.012	1165.0	0.000
25	0.326	-0.182	1181.4	0.000
26	0.304	-0.082	1195.8	0.000
27	0.270	-0.076	1207.3	0.000
28	0.244	0.081	1216.8	0.000
29	0.226	0.004	1224.9	0.000
30	0.159	0.007	1229.1	0.000
31	0.197	0.134	1235.5	0.000
32	0.184	-0.064	1241.1	0.000
33	0.184	0.049	1246.8	0.000
34	0.183	0.038	1252.5	0.000
35	0.177	-0.063	1257.9	0.000
36	0.195	-0.088	1264.6	0.000

Figure 3: ACF and PACF

Based on Figure 2 and Figure 3, generally, there is adequate evidence to conclude that the series of the road accidents in Malaysia is not stationary. Figure 2 shows that the trend pattern of the road accidents in Malaysia over the years fluctuates around a constant level. Hence, a horizontal pattern does not occur and this type of series is not stationary. The number of road accidents in Malaysia that do not increase or decrease consistently would be considered a not horizontal pattern. In Figure 3, the ACF shows a very slow decaying movement from the first lag to the last lag. Moreover, the decaying pattern shown in the first until forth PACF lags which indicates that the data is not stationary.

To further investigate the nature of stationary of the initial data, Unit Root Test is used to determine the changes of the data. Table 1 shows the result of Unit Root Test of the road accidents in Malaysia data.

Table 1: The Results of Unit Root Test

		t-Statistics	p-value
Augmented Dickey-Fuller test statistic		-0.364829	0.9101
Test Critical Value	1%	-3.492523	
	5%	-2.889669	
	10%	-2.581313	

The hypothesis of this test is; H_0 : concludes that the series contains a unit root versus the H_1 : which conclude that the series has no unit root. Thus, in order for the data to be stationary, it needs to reject the null hypothesis and concluded that the data does not comprise a unit root.

p-ISSN 1675-7939; e-ISSN 2289-4934

Table 1 shows that the probability value is 0.9101 which is larger than all the three critical values given at 1%, 5% and 10% level. Therefore, it can be concluded that the number of road accidents in Malaysia is not stationary.

Since the data is not stationary, the first order differencing on the data needs to be managed in order to reach the stationary and therefore carry out the primary assumptions of Box-Jenkins methodology that the data must be stationary in order to develop the ARIMA model. Let the series in first difference be w_t such that, $w_t = x_t - x_{t-1}$ where x_t is the current value and x_{t-1} is the preceding value.

3.3 Model Development

The first order differencing was performed in order to furnish the original stationary series.

Table 2: The Results of Unit Root Test after First Order Differencing

		t-Statistics	Probability
Augmented Dickey-Fuller test statistic		-4.421830	0.0005
Test Critical Value	1%	-3.492523	
	5%	-2.888669	
	10%	-2.581313	

Table 2 shows the output for a unit root test of the series after the first order differencing. From the output, the probability value 0.0005 is lower than any of the critical values which indicates that the series does not have a unit root. At any significant level of 5%, the null hypothesis can be rejected. Hence, it can be concluded that the data is stationary. At this stage, the order of differencing is equal to 1, *d=1*.

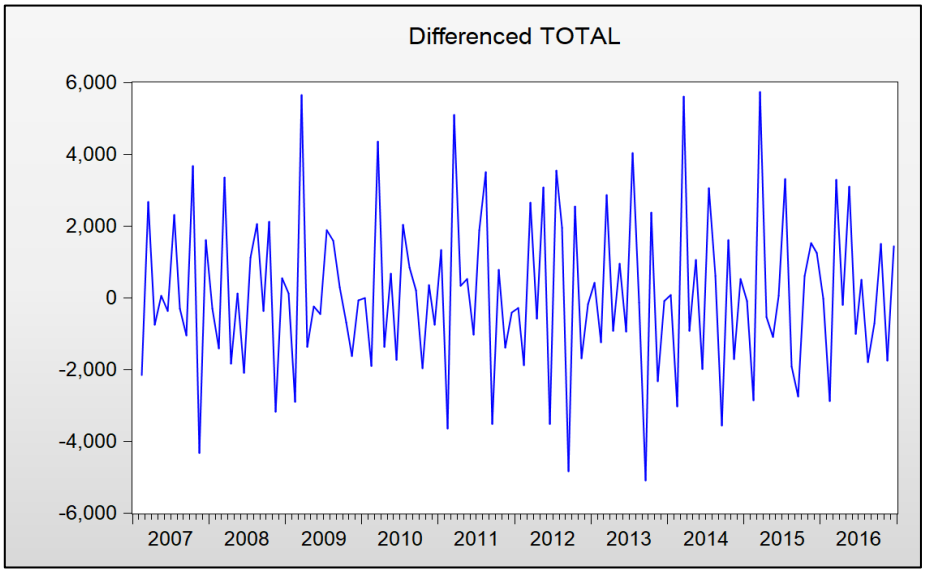
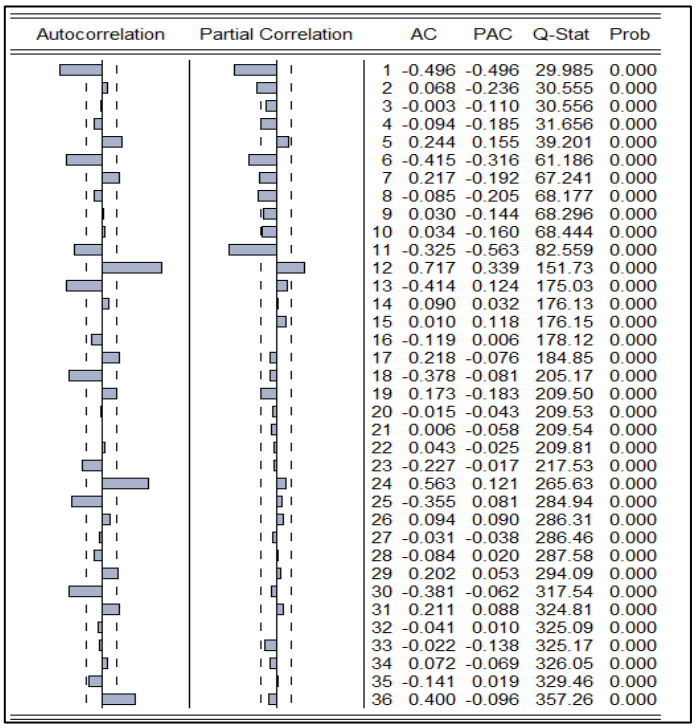


Figure 4: Graph after First Order Differencing

Figure 4 indicates the graph after the first order differencing. From the graph, it can be determined that it appears to be horizontal pattern. Hence, it can be concluded that the number of road accidents is stationary.

p-ISSN 1675-7939; e-ISSN 2289-4934

The ACF and PACF were also plotted to collect more conclusive evidence on its stationary condition (Figure 5). Both ACF and PACF do not show any noticeable decaying pattern thus adding more conclusive evidence on its stationary condition.



	AC	PAC	Q-Stat	Prob
1	-0.496	-0.496	29.985	0.000
2	0.068	-0.236	30.555	0.000
3	-0.003	-0.110	30.556	0.000
4	-0.094	-0.185	31.656	0.000
5	0.244	0.155	39.201	0.000
6	-0.415	-0.316	61.186	0.000
7	0.217	-0.192	67.241	0.000
8	-0.085	-0.205	68.177	0.000
9	0.030	-0.144	68.296	0.000
10	0.034	-0.160	68.444	0.000
11	-0.325	-0.563	82.559	0.000
12	0.717	0.339	151.73	0.000
13	-0.414	0.124	175.03	0.000
14	0.090	0.032	176.13	0.000
15	0.010	0.118	176.15	0.000
16	-0.119	0.006	178.12	0.000
17	0.218	-0.076	184.85	0.000
18	-0.378	-0.081	205.17	0.000
19	0.173	-0.183	209.50	0.000
20	-0.015	-0.043	209.53	0.000
21	0.006	-0.058	209.54	0.000
22	0.043	-0.025	209.81	0.000
23	-0.227	-0.017	217.53	0.000
24	0.563	0.121	265.63	0.000
25	-0.355	0.081	284.94	0.000
26	0.094	0.090	286.31	0.000
27	-0.031	-0.038	286.46	0.000
28	-0.084	0.020	287.58	0.000
29	0.202	0.053	294.09	0.000
30	-0.381	-0.062	317.54	0.000
31	0.211	0.088	324.81	0.000
32	-0.041	0.010	325.09	0.000
33	-0.022	-0.138	325.17	0.000
34	0.072	-0.069	326.05	0.000
35	-0.141	0.019	329.46	0.000
36	0.400	-0.096	357.26	0.000

Figure 5: ACF and PACF after First Order Differencing

3.4 Model Identification

Once the data is stationary and have fulfilled the principal assumption of application of the Box-Jenkins methodology, the next in this analysis is to perform the model identification. The process of identifying the suitable models to be fitted to the data series involved the analysis of the ACF and the PACF of the stationary series as depicted in Figure 5. In this study, both AR and MA have been denoted as *a* and *b* respectively. The investigation decided to choose and list down several models so that it is possible to make a comparison and select the best model out of all possible models for forecasting.

Based on Figure 5, both ACF and PACF diagrams show a few discernable spikes by pointing out the potential values of MA and AR. To identify the AR part of the model one needs to observed the PACF in Figure 5 from lag 1 to lag 11. One or possibly two significant spikes (exceeding the 2 standard error line) are observed, one at lag 1 and the other at lag 6 which represent the order of AR. Similarly, from the ACF diagram, two significant spikes are observed at lag 1 and 11. These two spikes can be used to specify the MA part of the model. The following models have been identified and estimated using E-views software.

- ARIMA (1,1,1)
- ARIMA (1,1,2)
- ARIMA (2,1,1)
- ARIMA (2,1,2)

p-ISSN 1675-7939; e-ISSN 2289-4934

3.5 Performance Evaluation

The following models are assessed in order to choose the best models for modelling road accidents in Malaysia. Based on Table 3, all parameters based on the probability value is statistically significant (probability value less than 0.05).

The probability value for AR(1) and MA(1) are statistically significant while that of MA(2) is not statistically significant since the probability value is more than 0.05 (Table 4). From Table 5 the parameter based on probability value is statistically significant.

Table 3: ARIMA (1,1,1) Model

Variable	Coefficient	Std. Error	t-Statistic	Probability value
AR (1)	0.994605	0.013431	74.05475	0.0000
MA (1)	-0.656947	0.066161	-9.929475	0.0000

Table 4: ARIMA (1,1,2) Model

Variable	Coefficient	Std. Error	t-Statistic	Probability value
AR (1)	0.994784	0.013217	75.26387	0.0000
MA (1)	-0.654711	0.094519	-6.926782	0.0000
MA (2)	-0.007091	0.124928	-0.056762	0.9548

Table 5: ARIMA (2,1,1) Model

Variable	Coefficient	Std. Error	t-Statistic	Probability value
AR (1)	0.012545	0.192076	5.271592	0.0000
AR (2)	-0.664545	0.092795	-7.161438	0.0000
MA (1)	-0.670802	0.138052	-4.859060	0.0000

Table 6: ARIMA (2,1,2) Model

Variable	Coefficient	Std. Error	t-Statistic	Probability value
AR (1)	0.217029	0.840361	0.258257	0.7967
AR (2)	0.774092	0.836560	0.925327	0.3568
MA (1)	0.140900	0.790360	0.178273	0.8588
MA (2)	-0.550053	0.480175	-1.145526	0.2544

All parameter in Table 6 are not statistically significant since the t-statistic value is less than 2 in absolute terms and the probability value is more than 0.05.

To determine which of the models fits the best, two criteria will be used, that is the AIC and BIC and the results are summarised in Table 7. Note that the E-view software does not provide the Box-Pierce (Ljung-Box) statistic. Since ARIMA (1,1,2) and ARIMA (2,1,2) are not significant, the parameters of ARIMA (1,1,1) and ARIMA (2,1,1) are compared.

p-ISSN 1675-7939; e-ISSN 2289-4934

Table 7: Model Estimation and Validation

MODEL	AIC	BIC
ARIMA (1,1,1)	**17.982**	**18.075**
ARIMA (2,1,1)	17.999	18.115

Based on Table 7, ARIMA (1,1,1) is the best model. ARIMA (1,1,1) has the smallest value of AIC and BIC as compared to ARIMA (2,1,1), which are 17.982 and 18.075 respectively. Furthermore, the concept of model's simplicity (parsimony) also points toward this model type.

4. CONCLUSION

Time series analysis of the data from the years 2007 until 2017 showed that the number of road accidents in Malaysia are increasing. From the model evaluations, it was found that ARIMA(1,1,1) was the best model since it presented the smallest values of AIC and BIC.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	36614.63	5327.038	6.873356	0.0000
AR(1)	0.994605	0.013431	74.05475	0.0000
MA(1)	-0.656947	0.066161	-9.929475	0.0000
SIGMASQ	3447303.	516665.4	6.672215	0.0000
R-squared	0.832582	Mean dependent var		36867.25
Adjusted R-squared	0.828253	S.D. dependent var		4556.761
S.E. of regression	1888.432	Akaike info criterion		17.98246
Sum squared resid	4.14E+08	Schwarz criterion		18.07538
Log likelihood	-1074.948	Hannan-Quinn criter.		18.02019
F-statistic	192.2926	Durbin-Watson stat		1.975908
Prob(F-statistic)	0.000000			

Figure 6: The Fitted Output

Figure 6 shows the fitted model for ARIMA(1,1,1). The data were a good fit since the coefficient of the determination, is 83.26%. Therefore, the fitted model for the number of road accidents in Malaysia can be written as:

$$\hat{y}_t = 36614.63 + 0.9946y_{t-1} + 0.6569e_{t-1} + e_t$$

This study would provide the best model for forecasting the road accidents based on the best fitted model that had been chosen. Therefore, the road authority and the agencies involved can take intervention and make a prediction about road accidents using the best-fitted model in this study. Other than that, it also can be beneficial for the road authorities in reducing the rate of accident cases in Malaysia.

REFERENCES

[1] Alias (2011). Introduction Business Forecasting- a practical approach. UiTM Press

[2] Avuglah, R. K., Adu-Poku, K. A., & Harris, E. (2014). Application of ARIMA Models to Road Traffic Accident Cases in Ghana. *International Journal of Statistics and Applications, 4*(5), 233-239. doi:10.5923/j.statistics.20140405.03

p-ISSN 1675-7939; e-ISSN 2289-4934

[3] Balagun, O. S., Oguntende, P. E., Akinrefon, A. A., & Modibbo, U. M. (2015). The Comparison of the Performance of ARIMA and MA Model Selection on Road Accident Data in Nigeria. *European Journal of Academic Essays 2(3): 13-31*

[4] Jabatan Siasatan dan Penguatkuasaan Trafik Bukit Aman (2018, March). Number of Road Accident in Malaysia from 2007-2016.

[5] Kumar, S. V., & Vanajakshi, L. (2015). Short-term traffic flow prediction using seasonal ARIMA model with limited input data. *European Transport Research Review, 7*(3). doi:10.1007/s12544-015-0170-8

[6] Mahmoud, A. Z. (2017). Forecast Car Accident in Saudi Arabia with ARIMA Models. *International Journal of Soft Computing and Engineering (IJSCE).* ISSN: 2231-2307, Volume-7 Issue-3, July 2017

[7] Rohayu, S., Sharifah Allyana, S., Jamilah, M., & SV, W. (2012). Predicting Malaysian Road Fatalities for Year 2020,MRR 06/2012,. *Kuala Lumpur: Malaysian Institute of Road Safety Research.*

[8] World Health Organization (2009), Global status report on road safety: time for action, Gevena: World Health Organization.

p-ISSN 1675-7939; e-ISSN 2289-4934