Box-Jenkins and State Space Model in Forecasting Malaysia Road Accident Cases

Wan Zakiyatussariroh Wan Husin\textsuperscript{1,*}, Adlina Sofia Afdzal\textsuperscript{1}, Nur Lisa Hashim Azmi\textsuperscript{1} and Siti Auni Taqiah Sheikh Hamadi\textsuperscript{1}

\textsuperscript{1}Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Kelantan, 18500 Machang Kelantan, Malaysia

*wanzh@uitm.edu.my

Abstract. Road accident is one of the main causes of death and injury worldwide in this fast-paced modern world. Many developing countries, including Malaysia, are facing serious road accident problems. Forecasting road accident cases has become an important step towards setting the road safety target. Hence, this study aims to develop forecasting models and forecast future trends of monthly road accident cases in Malaysia. The data set on monthly number of accident cases from January 2001 to December 2019 was provided by Polis Diraja Malaysia (PDRM). Box-Jenkins and State space models were developed using the data under study. The models were then evaluated based on in-sample and out-sample evaluation using lowest root mean square error, mean absolute percentage error and mean absolute error. The results show that the basic structural state space model with trend and seasonal component was the most appropriate model in forecasting road accident cases in Malaysia. The 10-year ahead forecast from January 2020 to December 2030 shows that monthly road accident cases in Malaysia have a constant inclining pattern for each year. It is hoped that the finding from this study could become a reference for the authorities of Malaysia in making recommendations in order to improve road safety and reduce road traffic accidents in Malaysia.

Keywords: Box-Jenkins Model, Forecasting, Local Level Model, Road Accident Cases, SARIMA Model, State Space Model

1. Introduction
Road accident is one of the main causes of death and injury worldwide in this fast-paced modern world. According to World Health Organization (WHO), almost one million people are killed each year, three million are severely disabled for life, and thirty million are injured in road accidents \cite{1}. Death caused by road accidents around the world has become a global upward trend in recent years. The lives of approximately 1.35 million people in the world are cut short every year as the result of road traffic crashes \cite{2}. There will be around two million deaths per year by 2030 and more than half of the victims of road fatalities are young adults between the ages of fifteen and forty-four \cite{3}. The problem of death as a result of road accidents is now recognised by every nation as a global phenomenon. In addition, between 20 and 50 million people have suffered non-fatal injuries with many incurring disabilities for their injury. Many developing countries, including Malaysia, are facing serious road accident problems. Relative to its population, Malaysia has one of the highest traffic accident rates in the world. Malaysia
has been ranked as the seventh highest country in the world for the total number of road accidents since 2012 to 2018. The year 2018 showed the most dangerous year for Malaysian roads especially during festive season with 548.6 thousand of road accident cases [4]. The total number of road accident cases has increased approximately 41 percent over the last decade, with 369,319 and 521,466 for the year 2007 and 2016, respectively [5]. As a result, the total number of deaths increased from 6,282 to 7,152. Moreover, Malaysia has had the highest fatalities per 100,000 population worldwide since 1996.

Road accidents may lead to injuries and fatalities affecting individuals, families and communities. It burdens the healthcare delivery system with the occupation of limited hospital beds and the utilization of resources, as well as the loss of productivity and income with social and economic consequences. Therefore, to prevent and reduce the recurrences of accidents in the future, effective safety approaches and remedial actions should be taken with the help of factual statistical data [6]. Forecasting road accident cases is an important step towards setting the road safety target [7]. According to [8], forecasting of road accidents is useful in monitoring the effectiveness of various road safety policies implemented to minimize the occurrence of accidents. Methods of forecasting can also improve the communication between scientists and policy makers and, ultimately, lead to better planning and decision-making. Usually, the predictions of road accident cases are useful in providing a better understanding of accident trends and the effectiveness of existing safety measures. It is, therefore, important for security planners to evaluate the current security policies and measures by looking at future accident trends and taking corrective actions [9]. Forecasting road accident cases helps traffic analysts to predict future trends in road accidents so that appropriate corrective changes can be made to the existing road safety policy [10]. Thus, research and forecasting should be done in order to save costs and reduce serious accidents in the future [11]. In addition, the finding from forecasting road accidents can be used to draw attention of governments and other stakeholders to the problem of road accidents and their consequences for public health and Sustainable Development Goals (SDG) in reducing road traffic. This is because, one of the SDGs is transforming this country into good health and a well-being [12]. To alleviate this issue, the development of appropriate forecasting models, which is able to produce good accident forecasts, is becoming an important agenda. Hence, a lot of studies have been conducted by previous researchers using various forecasting models [3, 8-10, 13-22]. In Malaysia, there are numerous studies related to forecasting road accidents; however, most of the studies focused more on forecasting road accident fatalities rather than road accident cases [16, 17, 18, 19, 20].

Currently, the most commonly used model for forecasting road accident cases is Box-Jenkins model. Several researchers used this model to forecast road accident cases [3,19 - 23]. Normally, Box-Jenkins model is very suitable for series data with stationary assumption which has been rarely fulfilled. The process of transformation by removing the element of trend and seasonal component has to be done in ensuring the stationarity of the series. In this case, important information might be lost if it goes through this process [24]. One of the approaches that may account for this problem would be state space model that are able to model the time series component such as seasonal and trend instead of removing the components through differencing transformation as in Box-Jenkins method. The state space modelling has been applied by previous researchers to a variety of time series problems such as in population, environmental, engineering, finance and many other study areas [24, 25, 26, 27, 28]. Besides, numerous researchers have used state space model in forecasting road accident cases [13, 29,30, 31]. According to [13], the method of state space is a superior alternative for traffic crash predictions. However, based on the researcher’s knowledge in the case of Malaysia, to date, there have been limited studies in forecasting road accident cases using state space model. One of the studies that applied univariate state space model was done by [31] which used local linear trend model in forecasting annual road accident cases. Another study was done by [24]; however, their aim focused more on investigating the contribution factors of road accidents as well as determining the effectiveness of structural time series in prediction, but not focusing on forecasting Malaysian road accident cases in the future. Hence, this study is conducted to forecast monthly road accident cases in Malaysia using state space model with the existence of trend and seasonal component. Then, the performance of the state space model is compared to the Box-Jenkins model.
2. Methodology

2.1. Data Description
The scope of this study concerned the forecast of road accident cases in Malaysia only. This study used secondary data. The data for this study were taken from 'Jabatan Siasatan and Penguatkuasaan Trafik PDRM Ibu Pejabat Polis Bukit Aman'. The dataset used involved road accident cases that covered the period of January 2001 to December 2019 (228 months, equivalent to 19 years). The data consisted of the number of road accident cases per month for each year. All datasets provided by PDRM covered all states in Malaysia. Since the data were monthly-based, the analysis may consist of trends, irregularities and seasonal components. Time plot was used to examine the pattern of monthly road accident cases. In addition, the trend analysis was also carried out in order to enable the researchers to understand the historical trend of datasets in monthly road accident cases.

2.2. Box-Jenkins Model
Box-Jenkins model refers to a systematic method in identifying, fitting, checking and using integrated autoregressive moving average (ARIMA) models. A Box-Jenkins model consists of an autoregressive model of order p (known as AR(p)) model, a moving average model of order q (known as MA(q)) model and combination of AR(p) and MA(q) which is known as ARMA(p,q) model. The classical Box-Jenkins models assume the time series is stationary, that is, the mean and variance of the series are essentially constant through time. When stationarity is not an issue, the ARMA (p,q) model can be written as follows:

\[ y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \]  \hspace{1cm} (1)

where \( y_t \) represents the time series data with stationary mean and variance, \( \phi_1, ..., \phi_p \) as the autoregressive parameters, \( \theta_1, ..., \theta_q \) are the moving average parameters, and \( \varepsilon_t, ..., \varepsilon_j \) are a series of unknown random errors (or residuals) assumed to follow a normal distribution [34]. In the case of series that are not stationary, the transformation, by taking the difference of the non-stationary time series values, is needed. In this case, the integrated ARMA(p,q) model needs to be considered. The general term for integrated ARMA(p,q) model is ARIMA (p,d,q) where ‘d’ is the number of time the variable needs to be differenced in order to achieve stationarity. The ARIMA (p,d,q) model can be written as follows:

\[ y'_t = \mu + \sum_{i=1}^{p} \phi_i y'_{t-i} + \varepsilon_t - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \]  \hspace{1cm} (2)

where \( y'_t \) is the difference of series while \( \{ \phi \}, \{ \theta \} \) and \( \{ \varepsilon \} \) are defined as previously. However, an ARIMA model only handles data with a trend component and it does not support time series with a seasonal component [32]. When both trend and seasonality are present, both non-seasonal difference and seasonal difference need to be applied. Then, an extension to ARIMA that supports the direct modelling of the seasonal component of the series is called seasonal ARIMA (SARIMA) which is denoted as SARIMA (p,d,q) (P,D,Q)s with \( p \) as non-seasonal AR order, \( d \) as non-seasonal MA order, \( P \) as seasonal AR order, \( D \) as seasonal differencing, \( q \) as non-seasonal MA order, \( P \) as seasonal AR order, \( D \) as seasonal differencing, \( Q \) as seasonal MA order, and \( s \) as time span of repeating seasonal pattern. Then, SARIMA (p,d,q) (P,D,Q)s model can be written as follows [32]:

\[ y'_t = \mu + \sum_{i=1}^{p} \phi_i y'_{t-i} + \sum_{i=1}^{p} \Phi_i y'_{t-1s} + \varepsilon_t - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{j=1}^{q} \Theta_j \varepsilon_{t-1s} \]  \hspace{1cm} (3)

where \( \{ y'_t \}, \{ \phi \}, \{ \theta \} \) and \( \{ \varepsilon \} \) are defined previously, while \( \{ \Phi \} \) and \( \{ \Theta \} \) are the seasonal counterpart [32].

Box-Jenkins methodology adopted in this study consists of several iterative steps. The step begins with simple data investigation using a simple time plot in identifying the level of trend differencing (d) and seasonal differencing (D). Then, the series is transformed by differencing it in order to achieve the stationarity series. Augmented Dickey-Fuller (ADF) unit root test is also conducted to ensure the series is stationary. In addition, following Webel and Ollech [31], WO-test was used in checking the presence
of seasonality since the data under study is monthly data. Subsequently, model identification was done using time plots of the data, autocorrelations function (ACF), partial autocorrelations function (PACF) and other information. A class of simple ARIMA models was then identified. This amounts to estimating (or guesstimating) an appropriate value for $d$ followed by estimates for $p$ and $q$ (as well as $P$, $D$, and $Q$ in the seasonal time series setting). Then, AR and MA parameters were found via an optimization method like maximum likelihood in model estimation step. A diagnostic checking was done in which the possible fitted models was checked for inadequacies by studying the autocorrelations of the residual series (i.e., the time-ordered residuals). The autocorrelation test is called Box-Pierce Q-statistics or portmanteau test was used. The test statistics is:

$$Q = T \sum_{k=1}^{K} \hat{p}_k^2$$

(4)

where $\hat{p}_k^2$ is the $k$-th order sample autocorrelation of estimated residuals and $T$ is the sample size of the data [32]. The null hypothesis of uncorrelated residuals is rejected if the $Q$ exceeds the tabulated critical value or $p$-value less than level of significance. Finally, model selection criteria, which included the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), were used in the model selection. The formula to calculate AIC and BIC is as follows:

$$AIC = -2 \ln \ln (L) + 2g$$

(5)

$$BIC = -2 \ln \ln (L) + g \ln (n)$$

(6)

where $L$ is the value of the likelihood function, $g$ is the number of coefficients being estimated and $n$ is the number of observations [32]. The best model is the one with the smallest AIC and BIC value. If there is no model to be compared, the model selection uses the criterion of actual versus fitted. If the fitted value is not much different from the actual, it means the model is accurate. Thus, the model is good and suitable for forecasting purpose. These steps were then iterated until there appears to be minimal to no improvement in the fitted model.

2.3. State Space Model

State Space model is described by unobserved variables or parameters which are involved in problems of dynamic time series. Determining an unobserved series of vectors, $\{\alpha_1,...,\alpha_n\}$ is a state space modelling basic approach that assumes the development over time of a system under investigation in relation to an observed series of vectors, $\{y_1,...,y_n\}$. The relationship between the $\alpha_t$'s and the $y_t$'s is specified by the state-space model and the purpose of state-space analysis is to infer the relevant properties of the $\alpha_t$'s from the knowledge of the state-space model and the realisation of the observations $\{y_1,...,y_n\}$. It allows a general treatment on virtually any linear time series models through the general algorithms of the Kalman filter and the associated smoother. Furthermore, it permits the likelihood function to be computed. Univariate state space model for time series is also known as structural time series model. A structural time series model has a direct interpretation and is formulated straightforwardly as far as various imperceptibly or inactive components such as trend and seasonal.

The simplest state space model is the local level model. The model is written as:

$$y_t = \mu_t + v_t, \quad v_t \sim iid N(0, \sigma_v^2)$$

(7)

$$\mu_{t+1} = \mu_t + w_t, \quad w_t \sim iid N(0, \sigma_w^2)$$

(8)

where $y_t$ is observed as univariate series which consist of a trend component, $\mu_t$, and a noise component, $v_t$, where, $w_t \sim iid N(0, \sigma_w^2)$ is indepenent of $\{v_t\}$ and $E[v_t w_t] = 0$ [34]. This model is incomplete without the initial specification of $\mu_1$. $y_1$ is the observed asset for the sum of two unobserved components $\mu_1$ and $v_1$. The component $\mu_1$ is the state variable, $\mu_t$ represents fundamental value of the asset while $v_t$ is random deviations from the fundamental value. According to [34], in the local level model case, the state of the system at time $t$ is expressed by the dynamic properties relating to the state of the system at time $t+1$. 


The local linear trend model is the local level model with a slope which refers to trend component. The following is the local linear trend model;

\[
y_t = \mu_t + \nu_t, \quad \nu_t \sim iid \ N(0, \sigma^2_n)
\]

\[
\mu_{t+1} = \mu_t + a_t + w_t, \quad w_t \sim iid \ N(0, \sigma^2_w)
\]

\[
a_{t+1} = a_t + \epsilon_t, \quad \epsilon_t \sim iid \ N(0, \sigma^2_\epsilon)
\]

where \(a_t\) is the slope of the trend component. Here \(\nu_t, w_t\) and \(\epsilon_t\) are independent Gaussian white noise processes. The basic structural model is a local linear trend model with an additional seasonal component. In seasonal model, level or trend component and a seasonal component are composed of a structural in time series model and it can be written in Gaussian state space form as follows:

\[
y_t = \mu_t + \gamma_t + \nu_t, \quad \nu_t \sim iid \ N(0, \sigma^2_n)
\]

\[
\mu_{t+1} = \mu_t + a_t + w_t, \quad w_t \sim iid \ N(0, \sigma^2_w)
\]

\[
a_{t+1} = a_t + \epsilon_t, \quad \epsilon_t \sim iid \ N(0, \sigma^2_\epsilon)
\]

\[
y_{t+1} = -\sum_{j=1}^{5} y_{t+1-j} + \omega_t, \quad \omega_t \sim iid \ N(0, \sigma^2_\omega)
\]

where \(\mu_t\) is slowly changing component for trend, \(\gamma_t\) is periodic component of seasonal while \(\nu_t\) is irregular component (disturbance). The disturbances of \(\nu_t, w_t\) and \(\epsilon_t\) are independent for all \(t\). Then, \(\omega_t\) in seasonal equation is the sum of the seasonal effect while the number of ‘seasons’ is denoted as \(s\) [35].

Estimation in State Space model involves estimation of the unobservable state and unknown parameter. An analysis that involves the state space model would produce the underlying unobserved signal \(x_t\), given that \(y_s = \{y_1, \ldots, y_s\}\), to time \(s\) and \(P^s_t\) is for convenience which fulfil the Kalman filter and smoother as follows:

\[
x^s_t = E(x_t \mid y_s)
\]

\[
P^s_t = E((x_t - x^s_t)(x_t - x^s_t)^T)
\]

where \(P^s_t\) corresponds to mean squared error, \(E\) is the projection operator and \(T\) is the transpose of the \((x_t - x^s_t)\). An iterative filter can be estimated by the value of the state component. A successive one step ahead prediction conditional on the past and concurrent observations is produced by Kalman filter for the set variable observation, \(\{y_1, \ldots, y_t\}\) [34]. To estimate Kalman filter that is denoted in \(a_{t+1}^i\) for the state of point \(t + 1\) such that \(a_{t+1} = E(x_{t+1} \mid y_t)\) and \(P_{t+1} = var(a_{t+1} \mid y_t)\), it is represented in variance estimation of filtered state variable which becomes equation (12).

\[
a_{t+1} = a_t + K_t(y_t - F_t^a a_t)
\]

The value of \(K_t\) is called Kalman gain and \(F_t\) is fixed matrices for the respective coefficients. The smoothing algorithm focuses on the estimate of \(a_t\) that relates to the observation that realises during the same time period, \(y_t\). For an application to the filter state, smoothing algorithm and original Kalman filter could be specified in a similar way.

\[
a_{t-1} = a_t + J_{t-1}(a_t^\xi - a_t)
\]

where \(a_t^\xi\) represents smoothed estimates and \(a_t\) represents filtered estimate. The ratio between the variance of \((a_t^\xi - a_t)\) and the variance in \(a_t\) would be determined in the value of \(J_{t-1}\). Kalman filter can simply continue in forecasting a time series. At the end of the sample, the filter state at time point \(t = n\) is updated in equation;

\[
a_n = a_{n-1} + K_{n-1}(y_{n-1} - z'_{n-1}a_{n-1})
\]
For the last observation, $y_n$, the filter state at time point is $t = n + 1$ and is updated as follows:

$$a_{n+1} = a_{n-1} + K_{n-1} (y_n - z_{n-1}^T a_n)$$  \hspace{1cm} (21)

Next is the estimation of the unknown parameters of the model. In the initial mean $\mu_0$ and covariance $\Sigma_0$, the vector of unknown parameter is presented as $\Theta$, the transition matrix is $\Phi$ while the state and observation covariance matrices are $Q$ and $R$ with the input coefficient matrices of $\Gamma$ and $\Gamma'$ [34]. Since the assumption of initial state is normal $x_0 \sim N_p(\mu_0, \Sigma_0)$ with the errors are normal $w_t \sim iid N_p(0, Q)$ and $v_t \sim iid N_q(0, R)$, maximum likelihood is used. The likelihood of $L(\Theta)$ may be written as,

$$- \ln L(y(\Theta)) = \frac{1}{2} \sum_{t=1}^{n} \ln |\Sigma_t(\Theta)| + \frac{1}{2} \sum_{t=1}^{n} \epsilon_t(\Theta)' \Sigma_t(\Theta)^{-1} \epsilon_t(\Theta)$$  \hspace{1cm} (22)

where it is highly nonlinear and is a complicated function of the unknown parameters. The $- \ln L(y(\Theta))$ is the function to obtain a new set of estimates, say $\Theta^{(1)}$, so $\Sigma_t(\Theta)$ represents the covariance matrices of unknown parameter, $\Theta$, and $\epsilon_t$ is set of innovation.

2.4. Model Evaluation

At the initial stage of analysis, monthly data of accident cases were separated into two parts, which were training data (in-sample) from January 2001 to February 2016 (182 months) and testing data (out-sample) set from March 2016 to December 2019 (46 months). The evaluation procedure involved fitting the models using in-sample data and the models were then evaluated using out-sample data. The evaluation was conducted based on forecast accuracy using root mean square error (RMSE), mean absolute percentage error (MAPE) and mean absolute error (MAE). The model with minimum error was identified as the best performed model. The equations of these error measures are as follows:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} \epsilon_t^2}{n}}$$  \hspace{1cm} (23)

$$MAPE = \frac{\sum_{t=1}^{n} |\frac{\epsilon_t}{\gamma_t} \times 100|}{n}$$  \hspace{1cm} (24)

$$MAE = \frac{\sum_{t=1}^{n} \epsilon_t}{n}$$  \hspace{1cm} (25)

where $\epsilon_t = y_t - \hat{y}_t$, in which $y_t$ is the observed number of road accident cases at time $t$ and $\hat{y}_t$ is the forecasted number of road accident cases at time $t$ [32].

2.5. Forecasting

Forecasts of monthly road accident cases in Malaysia were generated for the next 10 years which is equivalent to 120 months from January 2020 to December 2030. This forecast period is selected based on the goal set by SDG with the target that within the year 2020 to 2030, road accident cases will be reduced by half [36]. Thus, the achievement of Malaysia towards this goal can be evaluated based on the forecast values generated from this study.

3. Results

This section presents the results in developing Box-Jenkins and state space model for road accident cases in Malaysia. Data from January 2001 to February 2016 (182 months) were used for in-sample evaluation while the remaining years of March 2016 to December 2019 (46 months) were used for out-sample evaluation. Then, the best model was used in forecasting monthly road accident cases for the next 10 years (equivalent to 120 months). The process of model development began by investigating trend pattern of road accident cases in Malaysia from January 2001 to December 2019. Trend plot was constructed to explain the basic pattern of road accident cases in Malaysia; thus, to discern any unusual observations or characteristics that exist. Figure 1 shows the trend pattern of monthly road accident cases.
in Malaysia from January 2001 to December 2019. The plot indicated that the number of road accident cases in Malaysia increased from 22,248 in January 2001 to 47,603 in December 2019 with the highest number of accident cases occurred during the months of festive seasons such as Hari Raya Aidil Fitri, Deepavali and Chinese New Year. This scenario indicated the existence of trend and seasonal components in the data set. From the year 2001 to 2002, the highest cases were in the periods of Christmas holidays and school holidays. Then, in 2003, 2004 and 2005, the highest cases were consistently registered in October. The reason for that could be due to the start of Ramadhan. For 2006, 2007, 2008 and 2009, the cases were also the highest in the month of October due to Hari Raya Aidil Fitri and Deepavali. Furthermore, in 2010, 2011, 2012 and 2013, there were Ramadhan and Raya events in the month of August, which may be the explanation for the cases to be high in that particular year, and finally, there was Thaipusam holiday in 2014.

The model development began by developing the ARIMA model. As for the first step, a simple data investigation was performed into the road accident cases data in Malaysia in order to understand the basic pattern of the series, thus, to recognise any presence of unusual observations. As observed in Figure 1, there was an obvious trend and seasonal pattern in the data which indicated that the data were non-stationary series with seasonal component. In order to gather further evidence of its stationary state, autocorrelation function (ACF) and partial autocorrelation function (PACF) in figure 2 were plotted. The ACF and PACF plots proved that the data were non-stationary with some seasonality since there were significant seasonal effects at lag 12, 24 and 36 as in ACF, while the PACF plot showed a significant spike after lag 12. In addition, the ACF plot decayed very slowly in road accident cases data, thus, the series was not stationary.

Augmented Dickey-Fuller (ADF) test was conducted to further confirm the presence of unit root. The test indicated that the series of road accident cases in Malaysia was non-stationary since ADF test statistic value of -0.155636 gave the probability value of 0.9405 which was significantly greater than
5% level of significance. Then, WO-test was conducted to confirm the existence of seasonal effect in the data. The WO-test result combined the result of QS-test, QSR-test and Kwman-test. Since the P-value of QS-test and QSR-test was <0.01 and Kwman-test was <0.002; hence, WO-test identified that road accident cases in Malaysia had the effect of seasonality. Since the data were non-stationary with some seasonality, Seasonal Mixed Autoregressive Integrated Moving Average (SARIMA) model was chosen to analyse the data. Since the data were clearly non-stationary with some seasonality, seasonal differencing at lag 12, 24 and 36 was generated as the data under study were monthly data. At this stage, the order of seasonal differencing was equal to 1, \( D=1 \). As presented in figure 3(a), after seasonal differencing, the data were still non-stationary as ACF plot still showed a wave pattern. Hence, the additional first differencing (at non seasonal level) was performed as shown in figure 3(b). It clearly shows that there was no particular pattern in the ACF and PACF plots after the first differencing at non seasonal level (Figure 3(b)). Thus, it can be concluded that the series of monthly road accident cases in Malaysia was stationary. At this stage, the order of differencing at non-seasonal level was equal to 1, \( d=1 \). Based on figure 3(b), there were significant spikes of the non-seasonal level in the ACF and PACF, respectively. This shows that the model contained AR and MA. The ACF plot contained lesser significant spikes compared to PACF plot. In addition, the significant spikes were only shown in lags 12 but none in lags 24 and 36 in ACF plots of seasonal level. It can be concluded that this is a seasonal MA model. In addition, there were significant spikes found in lag 12 and 24 of PACF plot which indicated that the data had seasonality effect.

Again, the ADF test results for difference series in figure 3(b) showed that the test statistic of 7.799206 had the p-value of < 0.0001 which was smaller than 5% level of significance. Thus, it can be concluded that the difference series of monthly road accident cases in Malaysia was stationary. In this study, model simplicity was applied. As shown in figure 3(b), it can be concluded that the Box-Jenkins model for this study was SARIMA(p,d,q \( )/P,D,Q\) \( m \) model with \( m = 12 \) as the data used were in monthly form. The early lags of 12, which included lags that were lesser than \( m \), were considered in order to identify the non-seasonal model. In this case, the significant spikes were spotted at lag 1 in both ACF and PACF plots of figure 3(b). Thus, the model consisted of AR(1), MA(1) and \( d=1 \). This made the possible non-seasonal model become ARIMA(1,1,1), ARIMA(0,1,1) and ARIMA(1,1,0). At the seasonal level, the lags to be focused were lags 12, 24 and 36 with the significant spikes only spotted at lag 12 after seasonal differencing. Therefore, the possible models were SARIMA(1,1,0)(0,1,1)\( _{12} \), SARIMA(0,1,1)(0,1,1)\( _{12} \) and SARIMA(1,1,1)(0,1,1)\( _{12} \). For model selection of Box-Jenkins, the AIC and BIC values were used to determine the best model of Box-Jenkins. As can be seen in table 1, the AIC
and BIC values for SARIMA(0,1,1)(0,1,1)_{12} was the lowest compared to other models. Therefore, it can be concluded that SARIMA(0,1,1)(0,1,1)_{12} is the best model for Box-Jenkins model of this study. Figure 4 shows the actual (black coded) and fitted (red coded) values for model SARIMA(0,1,1)(0,1,1)_{12} SARIMA(1,1,0)(0,1,1)_{12} and SARIMA(1,1,1)(0,1,1)_{12}, respectively. It can be seen that the fitted values for model SARIMA(0,1,1)(0,1,1)_{12} were very close to the actual values of accident cases compared to SARIMA(1,1,0)(0,1,1)_{12} and SARIMA(1,1,1)(0,1,1)_{12}. Hence, SARIMA(0,1,1)(0,1,1)_{12} is able to produce the most accurate fit and this model can be used for forecasting purposes.

| Model                | AIC     | BIC     |
|----------------------|---------|---------|
| SARIMA(1,1,0)(0,1,1)_{12} | 3704.76 | 3718.24 |
| SARIMA(0,1,1)(0,1,1)_{12} | **3668.83** | **3678.94** |
| SARIMA(1,1,1)(0,1,1)_{12} | 3750.29 | 3757.03 |

**Figure 4.** Actual and fitted trend of road accident cases in Malaysia based on SARIMA model

Next, the state space model was developed for road accident cases in Malaysia. Figure 5 is a decomposition of additive time series which were composed of seasonal time series by estimating the observed trend, seasonal and random components of road accident cases in Malaysia. The plot shows the original time series (top), the estimated trend component (second from top), the estimated seasonal component (third from top), and the estimated irregular component (bottom). Based on observed and trend components in figure 5, it showed an increasing value in the series which indicated a rise in the number of road accident cases in Malaysia. It was also clear from the decomposition plot of the existence of seasonal component. A seasonal effect was present since it showed a regular repetitive pattern while the random (irregular) component analysed the mean squared size (variance) of the model. According to [37], seasonality occurs when a seasonal pattern exists in the decomposition time series and there is
no seasonality effect in the model if the decomposition time series shows a horizontal straight line in which the value is equal to zero in seasonal component.

![Decomposition of additive time series for road accident cases in Malaysia](image)

**Figure 5.** Decomposition of additive time series for road accident cases in Malaysia

In this paper, a maximum likelihood was used in estimating the parameter of the state space model and the result is shown in table 2. The level disturbance of accident cases, $\sigma^2_w$ was $1.000e-01$ while the variance of the slope disturbance, $\sigma^2_\theta$ was $1.000e-03$. Next, $1.847e+06$ was the variance of observational disturbance, $\sigma^2_\epsilon$ and $6.936e+04$ was the variance value for seasonal disturbance, $\sigma^2_\omega$. The state level and slope component’s variance were close to zero. This indicates that the slope was best held constant and hardly changed over time while the seasonal component’s variance indicates that the seasonal variation did not appear to be the same over time. In estimating the initial parameter, the maximum likelihood method was used. Then, Kalman filter was used to compute the likelihood of the given set of parameters. Filtering and smoothing are important where filtering is to understand the current values and situation while smoothing is needed to understand the past situation. Figure 6 presents the outcomes of the filtering and smoothing values that have been summarized. It showed that the lines of filtering and smoothing did not differ much from each other. Moreover, the Kalman smoother is considered a good estimate of the covariance of the state variables as it gives more accurate analysis of the actual values.

| Parameter   | Disturbance Variance |
|-------------|-----------------------|
| Level       | $1.000e-01$           |
| Slope       | $1.000e-03$           |
| Observational | $1.847e+06$       |
| Seasonal    | $6.936e+04$           |

**Table 2.** Parameter estimate of state space model
Then, the performance of SARIMA and state space model was measured based on the result of in-sample evaluation and out-sample validation. The model that produced the smallest RMSE, MAPE and MAE was chosen as the best model to forecast road accident cases in Malaysia. Table 3 presents the comparison for in-sample evaluation and out-sample validation between SARIMA\((0,1,1)(0,1,1)_{12}\) model and State Space model. It showed that the smallest value for RMSE and MAE was State Space model while the smallest value for MAPE was SARIMA\((0,1,1)(0,1,1)_{12}\) model for in-sample evaluation. Based on the values of RMSE, MAPE and MAE, the State Space model was selected as the best model for out-sample evaluation part as it produced the smallest error for RMSE, MAPE and MAE. According to Leonard [38], the out-sample validation test has better goodness of fit rather than in-sample evaluation test in forecasting methods. This is because the out-sample evaluation test assesses better forecast accuracy. Therefore, it showed that the best model in forecasting road accident cases in Malaysia was State Space model. Then, this estimated model was used in forecasting the number of road accident cases in Malaysia from the year 2020 to 2030 in which the ten-step ahead forecast was generated.

### Table 3. Result of in-sample evaluation and out-sample validation

| Model                      | In-sample Evaluation (January 2001 – February 2016) | Out-sample Validation (March 2016 – December 2019) |
|----------------------------|-----------------------------------------------------|---------------------------------------------------|
|                            | RMSE       | MAPE   | MAE   | RMSE   | MAPE   | MAE   |
| SARIMA\((0,1,1)(0,1,1)_{12}\) | 1056.924  | 2.425  | 1056.924 | 1792.044 | 3.267  | 1476.017 |
| State Space                | 1033.835  | 2.450  | 794.022 | 1653.526 | 2.845  | 1294.068 |

Figure 7 shows the forecast values of monthly road accident cases from the year 2020 to 2030. There was an increase in the forecast value of road accident cases in Malaysia from the year 2020 to 2030. It can also be seen that the forecast values showed a consistent pattern for every month of each year. It showed the actual trend of accident cases from the year 2001 to 2019 and the future trend of ten years ahead for the year 2020 to 2030. The trend pattern of the forecast values was quite consistent with the historical data in which the patterns of actual and forecast line series were quite close. The trend clearly showed that the number of road accident cases gradually increased along the years. The road accident cases are predicted to be 63464 cases for December 2030. Moreover, the forecast line shows an increment value year by year and it seems to have a constant seasonal pattern.
4. Conclusions
This study developed a univariate state space model with the existence of trend and seasonal variation in forecasting the future trend of monthly road accident cases in Malaysia. At the same time, this study also investigated the current trend pattern of monthly road accident cases in Malaysia. The trend pattern of monthly road accident cases in Malaysia shows a gradual increasing pattern from the year 2001 to 2019. The trend seasonality was identified as it went up and declined in specific month each year. Thus, this study shows that the trend had a seasonality pattern. For the year 2001 to 2002, the highest case was in the month of December. For the year 2004 to 2009, the highest case was in the month of October. Lastly, for the year 2010 to 2014, the highest accident case was in the month of August. The repetitive cases show seasonality variation in the trend pattern. Based on the behaviour of the historical data, it concludes that the basic structural state space model of local linear trend model with an additional seasonal component is the most appropriate model in forecasting road accident cases in Malaysia. The forecast monthly road accident cases were then generated for 10 years ahead which was equivalent to 120 months from January 2020 to December 2030. The forecast values showed a constant inclining pattern for each year. It can be concluded that road accident cases in Malaysia are predicted to increase if nothing is being done to reduce it. However, the data coverage for this study was based on the historical data of accident cases from January 2001 to December 2019 which was before the pandemic of COVID-19. The trend of accident cases may be affected due to this unexpected pandemic; hence, future researcher may consider the data coverage until 2021 to consider the current situation of accident cases in Malaysia. This is to ensure less error, increase data accuracy and provide better analysis. In addition, it is also recommended for other researchers to develop calendar variation model because it seems that the seasonal pattern occurrence for road accident cases is inconsistent due to the festive calendar. This is due to the seasonal month of the road accident cases which is inconsistent for every year. It depends on the calendar event; for example, ‘Hari Raya’ or ‘Ramadhan’ which follows the Hijrah calendar. It is a possibility that a calendar variation model is able to produce a more accurate forecast for road accident cases in Malaysia.

Acknowledgments
The authors would like to thank Universiti Teknologi MARA (UiTM) for the facilities provided to conduct this study and Polis Diraja Malaysia (PDRM) for providing the data on the number of accident cases. The authors are grateful to the reviewers for their thoughtful in-depth comments which have been very helpful in the revision of this paper.
References

[1] Kareem A 2003 Review of global menace of road accidents with special reference to Malaysia—a social perspective, *Malays J Med Sci.*** 10(2) 31–39

[2] World Health Organization (WHO) Global plan for the decade of action for road safety 2011-2020 [Internet] 2020 [cited 24 January 2021] Available from: https://www.who.int/roadsafety/decade_of_action/plan/en/

[3] Wai A H C, Seng S Y and Fei J L W 2019 Fatality involving road accidents in Malaysia: A comparison between three statistical models *Proc. 2019 2nd Int. Conf. on Mathematics and Statistics* 101-105

[4] Hirschmann R Number of road accidents in Malaysia 2012-2018 [Internet] 2020 [cited 20 November 2020]. Available from: https://www.statista.com/statistics/794875/number-of-road-accidents-malaysia/

[5] Musa M F, Hassan S A and Mashros N 2020 The impact of roadway conditions towards accident severity on federal roads in Malaysia *PLoS ONE* 15(7) 1-12

[6] Chong H Y and Low T S 2014 Accidents in Malaysia construction industry: statistical data and court cases *International Journal of Occupational Safety and Ergonomics* 20(3) 503-513

[7] Wittenberg P, Sever K, Knoth S, Sahin N and Bondarenko J 2013 Determining quantitative road safety targets by applying statistical prediction techniques and a multistage adjustment procedure *Accident Analysis & Prevention* 50 566-577

[8] Manikandan M, Prasad V, Mishra A K, Kondur R K and Newtonraj A 2018 Forecasting road traffic accident deaths in India using seasonal autoregressive integrated moving average model *International Journal of Community Medicine and Public Health* 5(9) 3962-68

[9] Zolala F, Haghdoot A A, Ahmadjouybari T, Safari A, Bahrampour A, Baneshi M R and Razzaghi A 2016 Forecasting the trend of traffic accident mortality in West Iran *Health Scope* 5(3)

[10] Al-Ghamdi A S 1995 Time series forecasts for traffic accidents, injuries, and fatalities in Saudi Arabia *Journal of King Saud University Engineering Sciences* 7(2) 199-217

[11] Al-Ghamdi A 2002 Using logistic regression to estimate the influence of accident factors on accident severity *Accident Analysis and Prevention* 34(6) 729-41

[12] Department of Statistics Malaysia (DOSM) Sustainable development goals (SDG) indicators Malaysia 2018 [Internet] 2018 [cited 26 January 2021] Available from: https://www.dosm.gov.my/v1/index.php?r=column/cthemeByCat&cat=474&bul_id=L21OMWJHT21XaW1OQVxBQXNyYjBGUT09&menu_id=WjJGK0Z5bTk1ZEiVT09yUW1tRG41ZZ09

[13] Dong C, Xie K, Sun X, Lyu M and Yue H 2019 Roadway traffic crash prediction using a state-space model based support vector regression approach *PlosONE* 14(4)

[14] Sunny C M, Nithya S, Sinshi K S, Vinodini V, KG A L, Anjana S and Manojkumar T K 2018 Forecasting of road accident in Kerala: a case study *2018 Int. Conf. on Data Science and Engineering (ICDSE) (IEEE)* 1-5

[15] Abdella G M, Al-Khalifa K N, Tayseer M A and Hamouda A M S, 2019 Modelling trends in road crash frequency in Qatar State *International Journal of Operational Research* 34(4) 507-23

[16] Yusria D 2017 A time series analysis of road traffic fatalities in Malaysia Doctoral dissertation University of Malaya

[17] Sarani R, Allyana S and Voon W S 2016 Malaysian road fatalities prediction for year 2020 *Journal of the Australasian College of Road Safety* 27(2) 18-22

[18] Radzuan N Q, Hassan M H A, Majeed A P, Kassim K A A, Musa R M, Razman M A M and Othman N A 2020 Forecasting road deaths in malaysia using support vector machine *IndECC2019* (Singapore Springer) 261-67

[19] Law T H, Radin R S U and Wong S V 2005 The Malaysian government’s road accident death reduction target for year 2010 *IATSS Research* 29(1) 43-49
[20] Sarani R, Rahim S A S M, Marjan J M and Voon W S 2012 Predicting Malaysian road fatalities for year 2020 Transp Res Board 12-42
[21] Yaacob W F W, Husin W Z W, Aziz N A and Nordin N I 2011 An intervention model of road accidents: the case of OPS Sikap intervention Journal of Applied Sciences 11(7) 1105-12
[22] Aida S M N, Haziq H H, Nurul S H, Isnewati A M and Haslinda A M 2018 Modelling road accidents in Malaysia ESTEEM Academic Journal 14 66-76
[23] Balogun O S, Oguntunde P E, Akinrefo A A and 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) 13-33
[24] Junus N W, Ismail M T, Arsad Z and Rahman R A 2017 Malaysia road accidents influences based on structural time series analysis Applied Mathematics and Information Sciences 11(4) 1029-39
[25] Abd Nasir N F, Muzaffar A N, Rahmat S N E, Husin, W Z W and Abidin N Z 2020 Forecasting Malaysia under-5 mortality using state space model J. Phys.: Conf. Ser. 1496 012001
[26] Husin W Z W, Zainol M S and Ramli N M 2016 Common factor model with multiple trends for forecasting short term mortality Engineering Letters 24(1)
[27] Sharma S, Elvira V, Chouzenoux E and Majumdar A 2021 Recurrent dictionary learning for state space models with an application in stock forecasting Neurocomputing 450 1-13
[28] Dabrowski J J, Rahman A, George A, Arnold S and McCulloch J 2018 State space models for forecasting water quality variables: an application in aquaculture prawn farming Proc. 24th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining 177-185
[29] Hermans E, Wets G and Bossche F D V 2006 Frequency and severity of Belgian road traffic accidents studied by state-space methods Bureau of Transportation Statistics 9(1) 63-76
[30] Antoniou C and Yannis G 2013 State-space based analysis and forecasting of macroscopic road safety trends in Greece Accident Analysis & Prevention 60 268–76
[31] Junus N W M and Ismail M T 2014 Predicting road accidents: Structural time series approach American Institute of Physic 1605(1) 816-821
[32] Hyndman R J and Athanasopoulos G 2018 Forecasting: principles and practice OTexts
[33] Ollech D and Webel K 2020 A random forest-based approach to identifying the most informative seasonality tests (No. 55/2020) Bundesbank Discussion Paper (Frankfurt: Deutsche Bundesbank)
[34] Shumway R H and Stoffer D S 2016 Time Series Analysis and Its Applications with R Examples 1960 289–381
[35] Koopman S J 2011 State Space Time Series Analysis Department of Econometrics VU University Amsterdam Tinbergen Institute 1-39
[36] World Health Organization 2018 Global status report on road safety 2018 1-20
[37] Brownlee J. 2017 How to Decompose Time Series Data into Trend and Seasonality [Internet] [cited 13 February 2021] Available from: https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/
[38] Leonard J T (2000). Out-of-sample tests of forecasting accuracy: an analysis and review International Journal of Forecasting 16 437-450