Accident Analysis and Prevention 112 (2018) 21–29

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Road traffic accidents prediction modelling: An analysis of Anambra State, Nigeria

Chukwutoo C. Ihueze, Uchendu O. Onwurah*

Department of Industrial and Production Engineering, Nnamdi Azikiwe University, Awka, Nigeria

ARTICLE INFO

Keywords:
Road traffic crashes
Time series analysis
ARIMA model
ARIMAX model
Forecasting
Anambra State

ABSTRACT

One of the major problems in the world today is the rate of road traffic crashes and deaths on our roads. Majority of these deaths occur in low-and-middle income countries including Nigeria. This study analyzed road traffic crashes in Anambra State, Nigeria with the intention of developing accurate predictive models for forecasting crash frequency in the State using autoregressive integrated moving average (ARIMA) and autoregressive integrated moving average with explanatory variables (ARIMAX) modelling techniques. The result showed that ARIMAX model outperformed the ARIMA (1,1,1) model generated when their performances were compared using the lower Bayesian information criterion, mean absolute percentage error, root mean square error; and higher coefficient of determination (R-Squared) as accuracy measures. The findings of this study reveal that incorporating human, vehicle and environmental related factors in time series analysis of crash dataset produces a more robust predictive model than solely using aggregated crash count. This study contributes to the body of knowledge on road traffic safety and provides an approach to forecasting using many human, vehicle and environmental factors. The recommendations made in this study if applied will help in reducing the number of road traffic crashes in Nigeria.

1. Introduction

One of the major problems in the world today is the rate of crashes and deaths on our roads. Each year, an estimated 1.24 million people are killed in road crashes and up to 20–50 million others injured, costing over US $500 billion worldwide (WHO, 2013). The cost of road traffic injuries is estimated to be between 1–2% gross national product in low-and-middle-income countries, which is over US $100 billion a year (Jacobs et al., 2000). The current trends show that if urgent action is not taken, road traffic injuries could be the seventh leading cause of death by the year 2030, and ninety percent of these deaths occur in low and middle-income countries (WHO, 2015). Not to mention the emotional trauma of losing loved ones, psychological impact on crash victims and permanent disability because of road traffic crashes.

The causes of road traffic crashes (RTC) are multi-factorial and involve interaction of a number of pre-crash factors that include human, vehicles and road environment (Haddon, 1980). Many studies have been conducted to investigate and understand the factors that are contributing to RTC in order to provide countermeasures. For instance, Ojo (2014) examined the factors contributing to road traffic crashes in Ekiti State, Nigeria using linear regression analysis and found that over speeding (speed violation), drivers distraction and dangerous overtaking contributed significantly to road crashes in the State. Ogunmodede et al. (2012) found that over speeding, wrong overtaking, bad roads, sudden mechanical defects, alcoholic intake, tyre burst and heavy rainfall contributed to increasing rate of motorcycle road crashes in Oyo State, Nigeria. Olawole (2016) studied the impact of weather (rainfall and temperature) on road traffic crashes in Ondo State, Nigeria between 2005 and 2012, and found that the correlations between road traffic crashes and elements of weather were generally low and never exceeded 0.41.

Haadi (2014) analyzed the factors that contributed to road crash severity in Ghana's Northern Region using binary logistic regression. The study found that overloading and obstruction were the two most significant factors contributing to road crash severity in Ghana. Also, brake failure has been found by previous studies to be one of the factors that has contributed to road traffic crashes in the developing countries (Oduro, 2012; Oluwole et al., 2015).

The results of most of the road traffic crashes studies in the developed countries have been effective in developing countermeasures that have helped to reduce road traffic crashes in the developed countries (Abdel-Aty, 2003; Noland and Quddus, 2004; Aguero-Valverde and Jovanis, 2006). But, the case has not been the same in some developing countries, more especially in Nigeria where road traffic crashes'

* Corresponding author.
E-mail addresses: cc.ihueze@unizik.edu.ng (C.C. Ihueze), debest2006@yahoo.com (U.O. Onwurah).

https://doi.org/10.1016/j.aap.2017.12.016
Received 26 July 2017; Received in revised form 21 November 2017; Accepted 21 December 2017

statistics still reveal a serious and growing problem with absolute fatality rate rising rapidly (Atubi and Gbadamosi, 2015). The factors that might account for the situation include differences in road users' behaviour, traffic mix, road quality design, unworthy road vehicles, inefficient enforcement of road safety laws, and most importantly, inadequate road crash predictive models that addressed crash contributing factors. This calls for more extensive research on road crashes in order to develop countermeasures and policies that will help to reduce road crashes in Nigeria and position her towards attainment of United Nations decade (2011–2020) of action on road safety.

In order to come up with effective countermeasures or address the contributing factors in Nigeria, there is a need to understand how traffic crashes will change and grow over time. One of the most effective methods of forecasting future occurrences in order to verify significance of certain variables is time series analysis (Mcleod and Vingilis, 2008). The time series analysis method mostly used in road safety research is autoregressive integrated moving average (ARIMA) models proposed by Box and Jenkins (1976). According to Quddus (2008), ARIMA model is the best crash predictive model for aggregated time series count data. Most of the past studies using ARIMA models used only the aggregated crash count without considering the explanatory variables that influenced the crash occurrences (Adu-poku et al., 2014; Avuglah et al., 2014; Balogun et al., 2015; Salifu, 2016; Sanusi et al., 2016).

However, some past studies have incorporated exogenous variables observed over the same period in investigating the relationships between the crash frequency or the severity and the contributing factors (Law et al., 2005; Quddus, 2008; Li and Chen, 2009; Bergel-Hayat et al., 2013; Theofilatos and Yannis, 2014). Although, some variables have been studied, few of these studies in the developed countries examined a comprehensive driver factors that were affecting road crashes. Therefore, more studies are still needed in incorporating broad spectrum of human, vehicle and environmental factors in ARIMA model for analysis and forecasting of RTC.

Hence, this study aims at developing an effective time series crash predictive model incorporating many human, vehicle and environmental (road) factors in comparison to solely using aggregated crash count. This study compares the performances of the univariate time series model (ARIMA model) and the multiple time series model (ARIMAX model) using Bayesian information criteria (BIC), mean absolute percentage error (MAPE), root mean square error (RMSE) and coefficient of determination (R-Squared) as accuracy measures. The model that performs better will be used for future crash analysis and forecasting in Anambra State, Nigeria.

The rest of the paper is organized as follows. Section 2 describes the data sources used for the analysis. Section 3 covers the methods used in data analysis. Section 4 presents the results and discussion of the findings. Section 5 presents the conclusion, recommendation and limitation of this study.

2. Data sources

In Nigeria, the Federal Road Safety Commission (FRSC) is the government agency with the statutory responsibilities for road safety administration. Among the various roles of FRSC are giving prompt attention and care to the victims of crashes, carrying out thorough investigation on the remote and immediate contributing factors to road crashes and filing their reports. They gather the crash information through on the spot assessment of crash scenes, vehicle, environmental conditions, and thorough interviews of the crash victims (drivers and passengers or pedestrians) and the onlookers. The records contain information on the types of road crashes, crash severity (fatal, serious and minor), categories of road users involved, vehicle type, number of vehicles involved, number injured, number killed, causes of crashes, date and time of occurrence, location among other things. The reports record crashes that accounted for at least one minor injury.

For this study, FRSC Anambra State Sector Command supplied data about road traffic crashes in the State for the period 1st January 2005 to 31st December 2015 (a total of 132 months). During this period, 1675 crashes were recorded; 18.84% involved minor injury crashes, 57.30% involved serious injury crashes and 23.86% involved fatal crashes. The 2005 and the 2006 crash data were not used because of some missing values in some months. In this study, an aggregated monthly count dataset of minor, serious and fatal crashes were used in the analysis. Only data from 2007 to 2015 were used. Fig. 1 shows the time series plot of the aggregated monthly count data of the number of crashes.

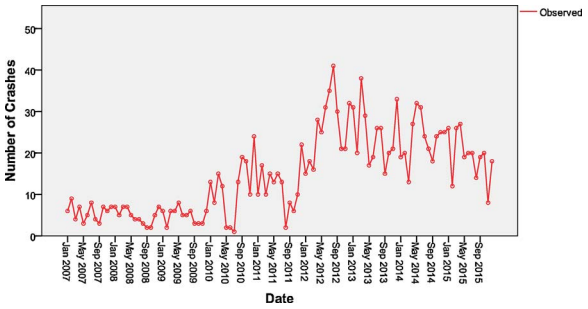


Fig. 1. Time Series Plot of Number of Crashes in Anambra State, Nigeria.

Each crash observation comprises a number of attributes (explanatory variables) relating to the victims of the crash, vehicle, and roadway and environmental conditions as judged by the investigating officers. The crash contributing factors observed over the same period and used in this study are tyre failure/burst, brake failure, over speeding (speed violation), loss of control, route violation, sign light violation, obstruction, wrong and dangerous overtaking, dangerous/reckless driving and weather condition (moderate and heavy rainfall only). The weather condition considered here is road crash that occurred under moderate and heavy rainfall. The descriptive statistics of the response variable and the explanatory variables are presented in Table 1. The crash prediction model that will be developed in this study using the dataset will also investigate the impact of each of the contributing factors on the monthly number of crashes.

Table 1
Descriptive Statistics of Road Traffic Crashes Data.

Variable	Number of months	Minimum	Maximum	Mean	Std. Deviation
Response Variable					
Number of Crashes	108	1	41	14.63	9.815
Explanatory Variables					
Over speeding	108	0	17	4.56	3.880
Tyre burst/ failure	108	0	5	0.80	1.092
Loss of control	108	0	17	3.00	3.496
Wrongful overtaking	108	0	6	0.57	1.061
Brake failure	108	0	8	1.69	1.921
Dangerous overtaking	108	0	5	0.67	0.917
Weather Condition	108	0	5	0.20	0.623
Route violation	108	0	8	0.86	1.300
Obstruction	108	0	3	0.30	0.600
Dangerous driving	108	0	11	2.85	2.335
Sign light violation	108	0	4	0.34	0.763

3. Methods of data analysis

The data collected were analyzed using ARIMA and ARIMAX models in SPSS version 22 software, and XLSTAT 2016 version was used for the Augmented Dickey–Fuller test in order to check for the presence of a unit root in the crash time series.

3.1. Autoregressive integrated moving average (ARIMA) model

ARIMA (p, d, q) model proposed by Box and Jenkins (1976) combines autoregressive(AR) and moving average(MA) models, and explicitly includes differencing in the formulation of the model suitable for univariate time series analysis. The AR model describes a time series in which the current observation depends on its preceding values, whereas MA model is used to describe a time series as a linear function of current and previous random errors. The general form of ARIMA (p, d, q) model is given as;

$$\nabla^d y_t = \frac{\theta(B)}{\varphi(B)}\varepsilon_t \tag{1}$$

Where y_t is the response series (monthly number of crashes), t is the time trend, ε_t is the random error (white noise)at a time period t, B is the backshift operator, ∇ represents the integrated processes (where $\nabla y_t = y_t - y_{t-1}$), d is the order of non-seasonal difference needed to achieve time series stationarity, and the other parameters in the model are defined as follows;

$$\theta(B) = (1-\theta_1 B-\theta_2 B^2-\ldots-\theta_q B^q) \tag{2}$$

$$\varphi(B) = (1-\varphi_1 B-\varphi_2 B^2-\ldots-\varphi_p B^p) \tag{3}$$

Where $\varphi_1, \varphi_2, \ldots, \varphi_p$ are the autoregressive (AR) parameters, $\theta_1, \theta_2, \ldots \theta_q$ are the moving average (MA) parameters, p is the order of autoregressive part and q is the order of the moving average part.

A statistical significant and adequate ARIMA (p, d, q) model for time series modelling and forecasting is formulated following Box-Jenkins methodology (Box and Jenkins, 1976). Box and Jenkins (1976) proposed a three-step iterative process of model identification, parameter estimation and diagnostic checking to determine the best parsimonious model.

3.1.1. Model identification

The first step in developing an ARIMA model is to determine if the time series is stationary or not. A stationary time series is one whose statistical properties such as mean, variance or autocorrelation are all constant over time. The stationarity or otherwise of the crash time series data used in this study was checked using the plot of autocorrelation functions and the Augmented Dickey–Fuller test. The Augmented Dickey–Fuller regression equation is given as in Eq. (4) (Dickey and Fuller, 1979).

$$y_t = \alpha + \rho y_{t-1} + \sum_{i=1}^{k} \varphi_i \Delta y_{t-i} + \beta t + \varepsilon_t \tag{4}$$

Where y_t represents the response variable (number of crashes), Δy_{t-i} is the time lagged change in the response variable, ε_t is the white noise error term, t is the time trend. In the Augmented Dickey–Fuller test, if the computed p-value is greater than the significance alpha value, one cannot reject the null hypothesis that says there is a unit root for the series. The presence of a unit root shows that the series is non-stationary, and it could be made stationary mostly by applying differencing. Once the stationarity is achieved, the next step is to determine the orders of the autoregressive (AR) and moving average (MA) terms using the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF).

3.1.2. Model parameter estimation

The maximum likelihood approach was used to estimate the parameters of the identified model and the t-values were used to check if the model generated is statistically significant or not. In this study, many ARIMA models were examined and the lowest Bayesian information criterion (BIC) was used to select the best model from the significant ARIMA models generated. The BIC is expressed as in Eq. (5) (Priestly, 1981).

$$BIC = n.\ln(RSS/n) + k.\ln(n) \tag{5}$$

Where n is the number of effective observations used to fit the model, k is the number of parameters in the model and RSS is the residual sum of square.

3.1.3. Model diagnostic checking

The adequacy of the model, considering the properties of the residuals, was checked using the residuals ACF and PACF, and the Ljung-Box statistics (Q^*). Q^* is obtained using Eq. (6) (Ljung and Box, 1978).

$$Q^* = n(n+2)\sum_{j=1}^{P} r_j^2/n-j \tag{6}$$

Where r_j is the residual autocorrelation at lag j, n = number of residuals, P = number of time lags in the test. If the p-value associated with the Q^* statistic is small (that is, $p < \alpha$), the model is inadequate. The model could be modified or a new one could be considered until a satisfactory model is determined.

3.2. Autoregressive integrated moving average model with explanatory variables (ARIMAX model)

ARIMAX model is a logical extension of pure ARIMA model that incorporates independent variables, which add explanatory value to the model. ARIMAX model is also referred to as Transfer Function model (Wei, 2006). The transfer function model for this study is a multiple-input, single-output transfer function model shown in Eq. (7) which is obtained by adding non-stationary input series (explanatory variables) to Eq. (1).

$$\nabla^d y_t = \sum_{k=1}^{K} v_k(B)\nabla^d X_{k,t} + \frac{\theta(B)}{\varphi(B)}\varepsilon_t \tag{7}$$

Where $\frac{\theta(B)}{\varphi(B)}\varepsilon_t$ represents the noise series of the system assumed to be independent of the explanatory variables, $x_{k,t}$ represents the crash explanatory variables (where k = 11), $v_i(B)$ is the transfer function for kth input series, and can be expressed as (Wei, 2006);

$$v(B) = \frac{w(B)B^b}{\delta(B)} \tag{8}$$

Where the numerator can be expanded to $\omega(B) = \omega_0-\omega_1 B-\ldots-\omega_s B^s$, and denominator can be expanded to $\delta(B) = 1-\delta_1 B-\ldots-\delta_r B^r$, s and r represent the orders of the polynomial, b is the delay parameter representing the actual time lag that lapses before the input series produces effect on the output series. In transfer functions, the same differencing applies to all the terms. Eq. (7) can be written in a more rational form for non-stationary input and output time series as;

$$\nabla^d y_t = \sum_{k=1}^{K} \frac{\omega_k(B)B^{bk}}{\delta_k(B)}\nabla^d X_{k,t} + \frac{\theta(B)}{\varphi(B)}\varepsilon_t \tag{9}$$

Building an ARIMAX (Transfer function model) is a similar iterative process as building a univariate Box-Jenkins ARIMA model, that is, model identification, parameters estimation and diagnostic checking.

3.2.1. Identification

Identification stage of transfer function involves prewhitening of both the input and output series, calculation of cross-correlation functions of the prewhitened series, and identification of order of b, s and r.

Prewhitening process was achieved by fitting ARIMA model for each input series sufficient to reduce the residuals to white noise; then, filtered the input series with the model to get white noise series. The same ARIMA model was used to filter the output series to get white noise residual series. The prewhitening process for non-stationary series is given by Eqs. (10)–(12);

$$\nabla^d x_{it} = \frac{\theta_x(B)}{\varphi_x(B)} e_{it} \tag{10}$$

$$e_{it} = \frac{\varphi_{x(B)}}{\theta_x(B)} \nabla^d x_{it} \tag{11}$$

$$\beta_{it} = \frac{\varphi_x(B)}{\theta_x(B)} \nabla^d y_{it} \tag{12}$$

Where e_{it} and β_{it} are white noise series with mean, zero and variance, σ^2.

The cross correlation functions between the prewhitened input series and the prewhitened output series were calculated at various lags, L (L = 0, ± 1, ± 2,... ± 7). The computed values were then compared to the theoretical impulse response functions of different orders in order to obtain some idea of the delay parameter b and the orders r and s of the transfer function between the output and input series (Law et al., 2005).

3.2.2. Estimation

The model parameters were estimated using maximum likelihood estimation method (Wei, 2006)

3.2.3. Diagnostic checking

The adequacy of the model was checked using the Ljung-Box statistics and the plots of ACF and PACF.

3.2.4. Multicollinearity analysis

The Collinearity between two or more explanatory variables was checked using tolerance and variance inflation factor (VIF). The tolerance can be computed using, $Tolerance = 1-R_j^2$ and variance inflation factor, $VIF_j = 1/tolerance = 1/1-R_j^2$. Where R_j is the coefficient of determination when x_j is regressed on all other predictor variables in the model.

4. Results and discussion

The purpose of this study is to develop accurate crash predictive model for future prediction of crash occurrences in Anambra State Nigeria using crash data from 2007 to 2015 using ARIMA and ARIMAX models. The crash dataset used in this study was divided into two parts. The first part (crash data from 2007 to 2014) was used to estimate the model parameters and the other part (2015 data) was used to validate the model using the estimated model parameters. The results obtained from ARIMA and ARIMAX models are presented in this section.

4.1. Result of ARIMA model

The time series plot of the number of crashes in Fig. 1 exhibits a systematic change, giving an evidence of trend in the data. This shows that the series is not stationary. Natural logarithmic transformation of the crash dataset was performed to stabilize the variance in the dataset. Even after the transformation, there was still trend in the series. Fig. 2 shows the sample autocorrelation function (ACF) of the transformed series, which obviously indicates that the series is not stationary since the autocorrelation coefficients at various lags are outside the confidence interval. The Augmented Dickey–Fuller test shown in Table 2 (the second column) confirmed that the series is non-stationary since the computed p-value (p = .444) is greater than the significance alpha level (α = 0.05). The test did not reject the null hypothesis that there is

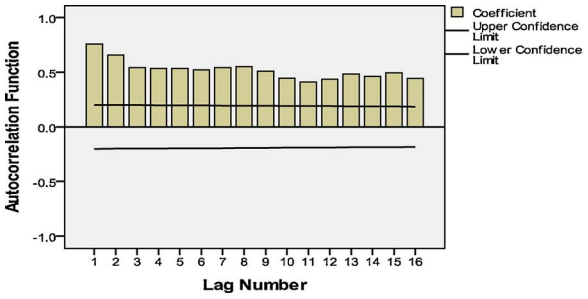


Fig. 2. Autocorrelation Function of the Number of Crashes.

Table 2
The Augmented Dickey–Fuller Test of the Number of Crashes.

Parameter	Original Time Series	Differenced Time Series
Tau (Observed value)	−2.251	−6.811
Tau (Critical value)	−0.801	−0.812
p-value (one-tailed)	.444	< .0001
Alpha	0.05	0.05

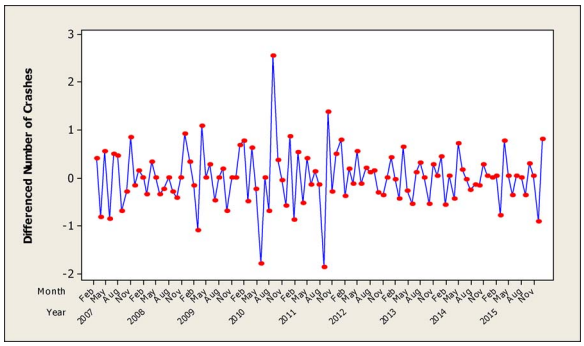


Fig. 3. Time Series Plots of the First Difference of the Number of Crashes.

a unit root in the series. This implies that at least first difference of the time series is needed to make it stationary.

Fig. 3 shows the time series plot of the first non-seasonal difference of the number of crashes, and it can be seen from Fig. 3 that the trend was removed from the time series after the first-order non-seasoning differencing was done. Fig. 4 shows the plot of ACF of the first difference of the crash time series. It can be seen that the graph of the ACF of the first difference of the time series decays very quickly, showing that the first-order non-seasonal difference of the series was enough to make the time series stationary. The ACF tails off after one significant spike at the first lag showing the presence of moving average component. Fig. 5 shows the partial autocorrelation function (PACF) of the first difference of the time series, which shows the presence of autoregressive component in the series. The stationarity of the first difference of the time series was also confirmed by the Augmented Dickey–Fuller test as can

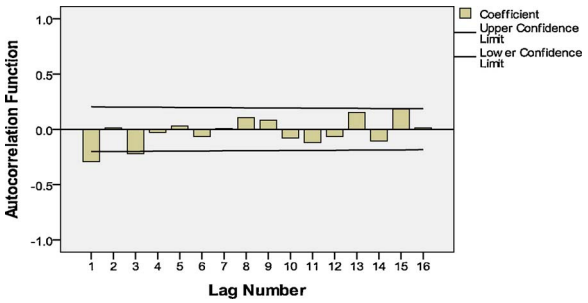


Fig. 4. Autocorrelation Function of the First Difference of the Number of Crashes.

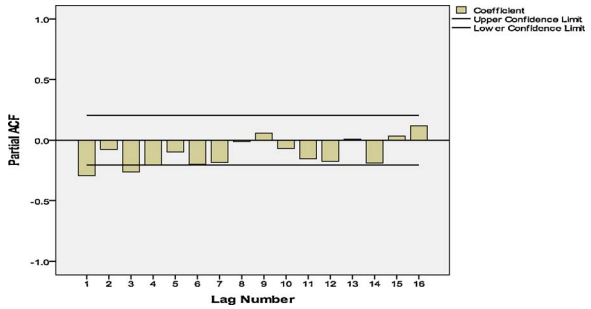


Fig. 5. Partial Autocorrelation Function of the First Difference of the Number of Crashes.

be seen in the third column in Table 2, where the computed p-value (p = < .0001) is less than the significant alpha level (α = 0.05).

From Figs. 4 and 5, it was not simple to identify the suitable ARIMA model for the number of crashes. To be certain of the best fit model for the RTC data, various ARIMA (p,d,q) models were examined at 95% confidence interval, their t-statistics were used to determine the significant models and the lowest Bayesian information criterion was used to determine the best model among the tentatively significant models. A significant model is the one, which all its parameters at 95% confidence interval are all significant. Table 3 shows the summary of the parameter estimation of various ARIMA models examined. From the table, ARIMA (1,1,1) has the lowest Bayesian information criterion of 3.655 and all its parameters are significant, hence, it was selected as the best model among the examined models. The model fit statistics in Table 4, look satisfactory, which shows that the ARIMA(1,1,1) model generated is fit to describe crash occurrences in the State. The R-squared value of 0.660 shows that over 66% of variations in the current aggregated monthly count dataset were explained by the past crashes and the past random shocks of the series. The positive coefficients of AR(1) and MA(1) show that there is a positive relationship between the observed aggregated number of crashes and, the time lagged observation and the lagged random shock, which implies that a unit increase in any of them, will increase the number of crashes while other parameters remain constant.

The diagnostic check to confirm the adequacy of ARIMA (1,1,1) model showed that the model is adequate for time series forecasting of road traffic crashes in Anambra State as can be seen in the plots of residuals ACF and PACF shown in Fig. 6. From the figure, the ACF and the PACF of the residuals are not significant at any lag, meaning that serial correlation is not significant between the error terms. Hence, the model is adequate. Also, from Table 4, the overall adequacy of the model checked using Ljung-Box (Q^*) statistic confirmed that the model is adequate and good fit for the aggregated monthly number of crash data in Anambra State, since the p-value (p = .380) computed is greater than the alpha value (α = 0.05).

The essence of fitting an ARIMA model is to properly understand the system and be able to make future predictions based on the historical pattern of the time series. The statistical significant and adequate ARIMA (1,1,1) model generated was used to make 12 months forecast of road crashes and the result is shown in Fig. 7. The red line shows the actual values (observed), the thin blue line shows the fit, and the out of the sample thick blue line shows the forecast made.

4.2. Result of ARIMAX model

The explanatory variables listed in Table 1 were incorporated into univariate ARIMA model to add explanatory values to the model. First order differencing was required to make each of the input series stationary. This was followed by prewhitening of each of the series using an adequate ARIMA model. The following ARIMA models were used in prewhitening the explanatory variables: ARIMA(0,1,1) for speed violation, ARIMA(2,1,2) for tyre burst, ARIMA(3,1,0) for loss of control, ARIMA(3,1,0) for brake failure, ARIMA(0,1,1) for dangerous overtaking, ARIMA(1,1,1) for dangerous driving, ARIMA(1,1,1) for route violation, ARIMA(0,1,1) for road obstruction, ARIMA(0,1,1) for poor weather, ARIMA(2,1,0) for sign light violation and ARIMA(1,1,1) for wrong overtaking were used in prewhitening the input series. Following the three-step iterations of model identification, parameter estimation and diagnostic checking, an ARIMAX model was generated for the number of crashes. The cross-correlation functions (CCF) between the prewhitened input series and the prewhitened output series in each case, as shown in Fig. 8, show a positive and significant correlation at lag 0. That is, there is no time delay (b = 0), which implies that the number of crashes increases instantly with increase in any of the explanatory variables.

Different values of the parameters ω and δ were estimated at different orders of s and r using SPSS version 22 software at 95% confidence interval and the t-statistics were used to determine the parameters that were significant. The non-significant parameters were removed from the model to reduce model complexity. The order with the lowest Bayesian information criterion was selected the best. The parameter estimates for the best ARIMAX model generated is shown in Table 5. All the parameters for both the input and the output series in the table are significant judging from their t-statistics, which are all

Table 3
Summary of Parameter Estimation for ARIMA Models.

Model	Parameter	Estimate	Standard Error	*t*-statistics	R-squared	Bayesian Information Criterion
***ARIMA(1,1,0)	AR 1	−0.289	0.099	−2.932	0.533	3.914
*ARIMA(2,1,0)	AR 1	−0.310	0.103	−3.000	0.548	3.940
	AR 2	–0.074	0.103	−0.716		
***ARIMA(0,1,1)	MA 1	0.515	0.089	5.799	0.605	3.747
***ARIMA(0,1,2)	MA 1	0.521	0.101	5.158	0.646	3.696
	MA 2	–0.233	0.101	2.316		
*ARIMA(0,1,3)	MA 1	0.437	0.101	4.327	0.655	3.728
	MA 2	0.074	0.110	0.671		
	MA 3	0.261	0.101	2.586		
***,[a]ARIMA(1,1,1)	AR 1	0.409	0.133	3.067	0.660	3.655
	MA 1	0.868	0.074	11.799		
*ARIMA(2,1,1)	AR 1	0.403	0.136	2.973	0.660	3.715
	AR 2	0.008	0.121	0.068		
	MA 1	0.866	0.088	9.828		
*ARIMA(1,1,2)	AR 1	0.417	0.334	1.248	0.660	3.715
	MA 1	0.879	0.352	2.499		
	MA 2	−0.011	0.259	−0.042		

* The non significant Models (*t*-statistics < 2).
*** The significant models (*t*-statistics ≥ 2).
[a] The selected model (the significant model with the lowest Bayesian information criterion).

Table 4
Model and Ljung-Box Statistics of ARIMA (1,1,1) Model.

Model	Model Fit Statistics				Ljung-Box Q (18)		
	R-Squared	Root Mean Square Error	Mean Absolute Percentage Error	Bayesian Information Criterion	Statistics	Degree of Freedom	p-value
ARIMA (1,1,1)	0.660	5.926	50.409	3.655	17.083	16	.380

greater than two in absolute term in each case.

All the explanatory variables have positive coefficients, which are significantly different from zero, showing that the increase in any of the contributing factors will lead to increase in the number of crashes. The goodness of fit of the model checked using the coefficient of determination (R-squared) as can be seen in Table 6, shows that over 97% of the variations in the number of crashes were explained by the 11 crash explanatory variables and the present and the lagged random shocks. This implies that the model is fit enough to describe crashes in the State. The findings of this study are consistent with the previous studies that used similar contributing factors in Nigeria or other developing countries. For instance, over speeding and dangerous overtaking (Ojo, 2014); tyre burst, wrong overtaking and heavy rainfall (Ogunmodede et al., 2012); obstruction (Haadi, 2014); and brake failure (Oduro, 2012).

Diagnostic Checking for the ARIMAX model selected was done to determine the adequacy of the model generated using the ACF and the PCF as shown in Fig. 9. The ACF and the PACF have only one significant spike at lag 14, but considering lags 1–13, the model is adequate. The overall adequacy of the ARIMAX model checked using the Ljung-Box Statistics is shown in Table 6. From the table, the Ljung-Box statistics, Q^{*} = 18.468 and its p-value is. 360. The p-value obtained is greater than the alpha value, α = .05, hence, the ARIMAX model generated is adequate and good fit for the time series data.

The results of multicollinearity analysis of the explanatory variables as shown in Table 7, show no Collinearity problem with the model since all the tolerances are more than 0.2 and all the variance inflation factors are less than 5. According to O'Brian (2007), a tolerance of less than 0.2 or 0.1 and/or VIF of 5 or 10 and above indicate Collinearity problem.

Since the model was found to be adequate and good fit for the crash data in the State, and also free from Collinearity problem, it was used to make 12 months forecast and is shown in Fig. 10. Before forecasting, 12 months forecast were made for each input series using the adequate ARIMA model for each input series (ARIMA models used during pre-whitening).

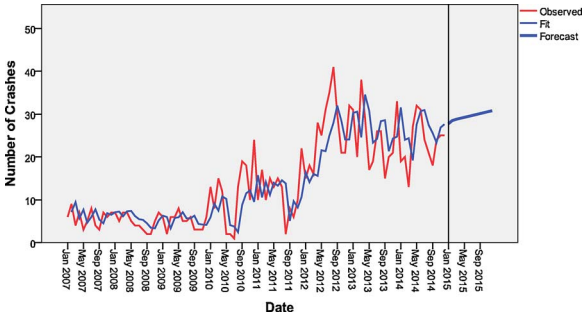


Fig. 7. ARIMA(1,1,1) Model Forecast for the Number of Crashes.

4.3. Comparison of the performances of the ARIMA model and the ARIMAX model

The performances of ARIMA and ARIMAX models examined showed that the ARIMAX model outperformed the ARIMA model as summarized in Table 8. The ARIMAX model has lower BIC (1.561), RMSE (1.466) and MAPE (12.254) than the ARIMA model with BIC (3.655), RMSE (5.926) and MAPE (50.409). It also has a better coefficient of determination (R-Squared = 0.979) than ARIMA model (R-Squared = 0.660), making it a more robust model for analysis of crash count data in the State.

The forecasts made by the two models were validated using 2015 crash data. Fig. 11 shows the comparison of the forecast made by each model to the 2015 actual crash data. The black line with circular marker shows the actual crash data, the red line with square marker shows the forecast made with the ARIMA model, and the green line shows the forecast made with the ARIMAX model. From the fig., the forecast made by the ARIMAX model is closer to the actual crash data than the one made by the ARIMA model. This implies that the ARIMAX model outperformed the ARIMA model. Although, the ARIMAX model performed better, there are still three months (February, August and November) their forecasted values are far away from the actual values. This suggests that the performances of other advanced forecasting techniques using the same crash data are needed to be explored.

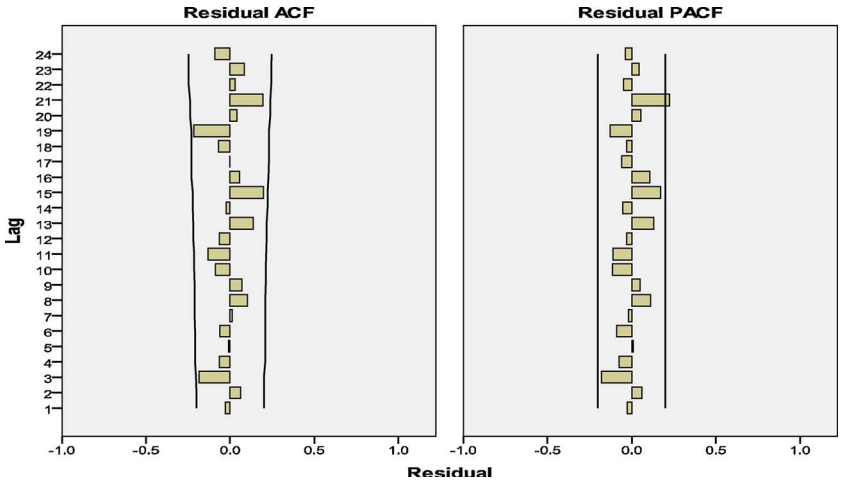
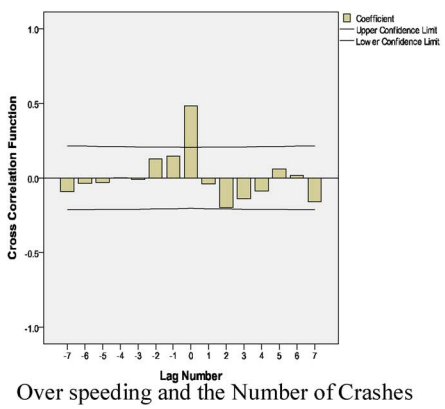
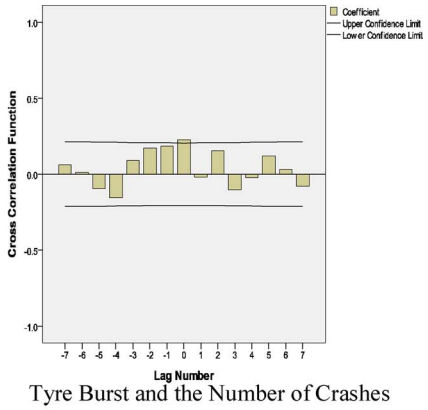


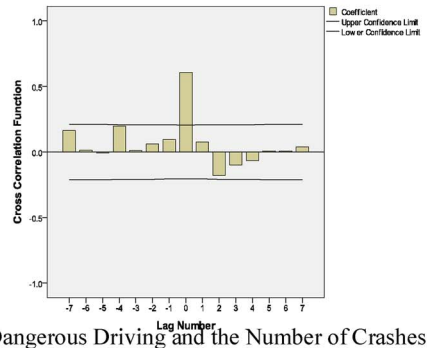
Fig. 6. Residuals ACF and PACF of ARIMA (1,1,1) Model for the Number of Crashes.



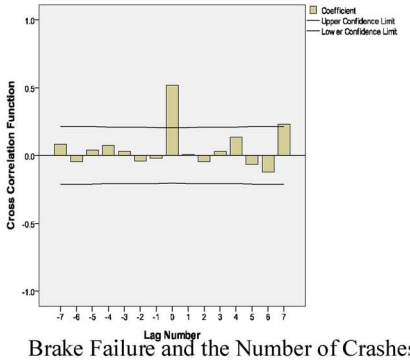
Over speeding and the Number of Crashes



Tyre Burst and the Number of Crashes



Dangerous Driving and the Number of Crashes



Brake Failure and the Number of Crashes

Fig. 8. Cross Correlation Functions of the Prewhitened Explanatory variables and the Number of Crashes.

Table 5
ARIMAX Model Parameters for the Number of Crashes.

Variable	Parameter	Lag	Estimate	Standard Error	t-statistic	p-value
Number of Crashes	Difference	1	0.994	0.449	2.216	.029
	MA	1				
Over speeding	Numerator	0	0.862	0.065	13.164	< .0001
Tyre failure/burst	Numerator	0	0.709	0.159	4.468	< .0001
Loss of Control	Numerator	0	0.936	0.064	14.605	< .0001
Wrong overtaking	Numerator	0	0.592	0.173	3.421	< 0.0001
Brake Failure	Numerator	0	1.086	0.115	9.418	< .0001
Dangerous overtaking	Numerator	0	0.929	0.243	3.830	< .0001
Weather Condition	Numerator	0	1.450	0.265	5.468	< 0.0001
Obstruction	Numerator	0	0.813	0.327	2.484	.015
Dangerous driving	Numerator	0	0.849	0.078	10.837	< .0001
Sign light violation	Numerator	0	0.580	0.211	2.755	.007
Route violation	Numerator	0	0.591	0.172	3.443	< .0001

Table 6
Model and Ljung-Box Statistics of ARIMAX Model for the Number of Crashes.

Model	Number of Predictors	Model Fit Statistics				Ljung-Box Q (18)		
		R-Squared	Root Mean Square Error	Mean Absolute Percentage Error	Bayesian Information Criterion	Statistics	Degree of Freedom	p-value
ARIMAX	11	0.979	1.561	12.030	1.466	18.468	17	.360

5. Conclusion

An analysis of road traffic crashes in Anambra State Nigeria has been carried out using Autoregressive integrated moving average (ARIMA) and Autoregressive integrated moving average with explanatory variables (ARIMAX) modelling approaches. The following were arrived at in this study: all the eleven contributing factors examined have significant effects on the number of crashes and an increase in any of them will increase crash occurrences in Anambra State as seen in their positive coefficients. Both the ARIMA (1,1,1) and the ARIMAX models are robust models that can be used to analyze and forecast road traffic crashes in the State. The performance of the ARIMAX model is better than that of the ARIMA model generated as can be seen from its lower BIC, RMSE and MAPE values; and higher coefficient of determination. Hence, ARIMAX model is more appropriate for analysis of road traffic crashes in the State. This implies that inclusion of a broad spectrum of human, vehicle and environmental related factors in RTC time series modelling and forecasting is necessary rather

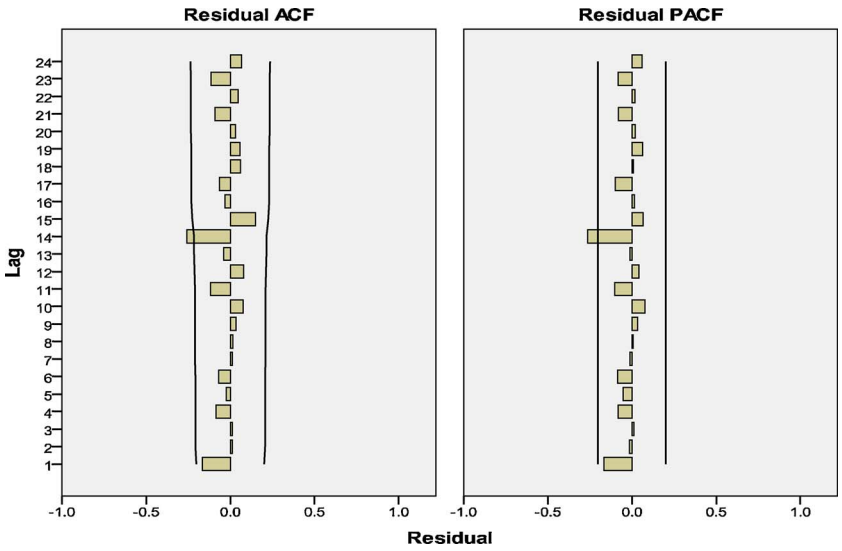


Fig. 9. Residuals ACF and PACF of the ARIMAX Model for the Number of Crashes.

Table 7
Result of Multicollinearity Analysis.

Explanatory Variable	Tolerance	Variance Inflation Factor
Over speeding	0.478	2.09
Tyre burst	0.809	1.23
Loss of control	0.456	2.17
Wrongful overtaking	0.787	1.27
Brake failure	0.517	1,93
Dangerous overtaking	0.547	1.82
Dangerous driving	0.636	1.57
Route violation	0.683	1.46
Obstruction	0.893	1.19
Poor weather condition	0.820	1.22
Sign light violation	0.905	1.11

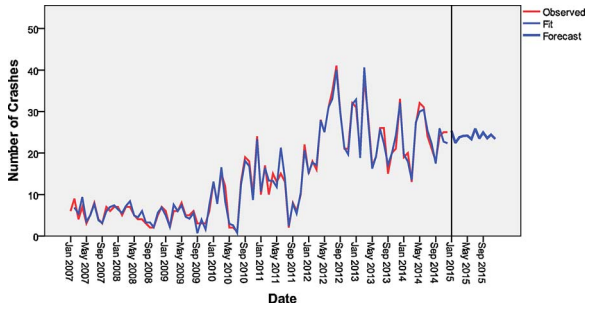


Fig. 10. ARIMAX Model Forecast for the Number of Crashes.

Table 8
Comparison of the Performances of the ARIMA and the ARIMAX Models.

Model	R-Square	Bayesian Information Criterion	Root Mean Square Error	Mean Absolute Percentage Error
ARIMA (1,1,1)	0.660	3.655	5.926	50.409
ARIMAX	0.979	1.466	1.561	12.254

than using solely aggregated crash count.

This study adds to the body of knowledge on road traffic safety and provides an approach to forecasting using many human, vehicle and environmental factors. The findings of this study would be useful to all stakeholders – road users, roadway designers and road construction companies, law enforcement agencies and policy makers in that they

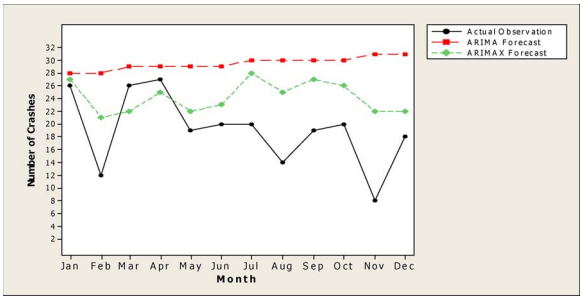


Fig. 11. Comparison of the 2015 Actual Crash Data to the Forecasted Values.

are in a better position to affect a variety of factors, hence, influencing road traffic safety. It would be useful in developing strategies that can help to prevent and reduce the number of crashes in Anambra State and by extension Nigeria. Most of the contributing factors that have significant effects on the number of crashes in this study have a lot to do with violation of road safety rules and regulation. This study recommends strict and total enforcement of all safety rules and policies in Nigeria, intensive safety awareness program to educate drivers and other road users, and installation of speed limiting devices in all the vehicles that ploy the roads in order to reduce incessant crashes because of over speeding. It also recommends introduction of road safety education in Nigeria School Curriculum from primary to tertiary institutions in order to inculcate safety consciousness in early stage of life.

One of the limitations of this study is that only the aggregated monthly crash count data from one State in Nigeria were used in the crash analysis and prediction, generalizing the forecast made with the model to the whole country is not advisable. But, considering the contributing factors used in the analysis, and the fact that the drivers' behaviour, road mix and environmental conditions are virtually the same in Nigeria, this study provides insight into crash contributing factors in Nigeria.

The future research effort in this regard should involve extending the research to the entire country and other developing countries; investigating the impact of other crash contributing factors such as use of mobile phone while driving, driving under the influence of alcohol and drugs, and demographic factors. Additionally, it would be useful to compare the prediction performances among other advanced modelling techniques such as artificial neural network.

References

Abdel-Aty, M., 2003. Analysis of driver injury severity levels at multiple locations using ordered probit models. J. Saf. Res. 34, 597–603.

Adu-Poku, K.A., Avuglah, R.K., Harris, E., 2014. Modeling road traffic fatality cases in Ghana. Math. Theor. Model. 4, 113–120.

Aguero-Valverde, J., Jovanis, P.P., 2006. Spatial analysis of fatal and injury crashes in Pennsylvania. Accid. Anal. Prev. 38, 618–625.

Atubi, A.O., Gbadamosi, K.T., 2015. Global positioning and socio-economic impact of road traffic accidents in Nigeria: matters arising. Am. Int. J. Contemp. Res. 5, 136–146.

Avuglah, R.S., Adu-poku, K.A., Harris, E., 2014. Application of ARIMA models to road traffic accident cases in Ghana. Int. J. Stat. Appl. 4, 233–239.

Balogun, O.S., Awoeyo, O.O., Akinrefon, A.A., Yami, A.M., 2015. On model selection of road accident data in Nigeria: a time series approach. Am. J. Res. Commun. 3, 139–177.

Bergel-Hayat, M., Constantinou, C., Yannis, G., 2013. Explaining the road accident risks: weather effects. Accid. Anal. Prev. 60, 456–465.

Box, G., Jenkins, G., 1976. Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

Dickey, D.A., Fuller, W.A., 1979. Distribution of the estimation for autoregressive time series with a unit root. J. Am. Stat. Assoc. 79, 355–367.

Haddon, W., 1980. Advances in the epidemiology of injuries as a basis for public policy. Public Health Rep. 95, 411–421.

Haadi, A., 2014. Identification of factors that cause severity of road accidents in Ghana: a case study of the Northern Region. Int. J. Appl. Sci. Technol. 4, 242–249.

Jacobs, G.D., Aaron-Thomas, A., Astrop, A., 2000. Estimating Global Road Facilities. Transport Research Laboratory (TRL) Report 445, Crow Thorne.

Law, T.H., Radin Umar, R.S., Wong, S.V., 2005. The Malaysian Government's road accident death reduction target for year 2010. IATSS Res. 29, 42–49.

Li, C., Chen, J., 2009. Traffic accident macro forecast based on ARIMAX model. Proceeding of the International Conference on Measuring Technology and Mechatronics Automation. IEEE Computer Society, Washington, DC, USA.

Ljung, G.M., Box, G.E.P., 1978. On a measure of lack of fit in time series models. Biometrika 65, 297–303.

Mcleod, A.I., Vingilis, E.R., 2008. Power computations in time series analysis for traffic safety interventions. Accid. Anal. Prev. 60, 1244–1248.

Noland, R.B., Quddus, M.A., 2004. Improvements in medical care and technology and reductions in traffic-related fatalities in Great Britain. Accid. Anal. Prev. 36, 103–113.

Priestley, M.B., 1981. Spectral Analysis and Time Series. Academic Press, London.

O'Brian, R.M., 2007. A caution regarding rules of thumb for variance inflation factors. Qual. Quant. 41, 673–690.

Oduro, S.D., 2012. Brake failure and its effect on road traffic accidents in Kumasi metropolis, Ghana. Int. J. Sci. Technol. 1, 448–453.

Ogunmodede, T.A., Adio, G., Ebijuwa, A.S., Oyetola, S.O., Akinola, J.O., 2012. Factors influencing high rate of commercial motorcycle accidents in Nigeria. Am. Int. J. Contemp. Res. 2, 130–140.

Ojo, A.L., 2014. Predominant causes of road traffic accident among commercial vehicle drivers in Ekiti State, Nigeria. Int. J. Hum. Soc. Stud. 2, 1–5.

Olawole, M.O., 2016. Impact of Weather on Road Traffic Accidents in Ondo State, Nigeria: 2005–2012, vol.1. Analele Universitatii din Oradea, Seria Geografie, pp. 44–53.

Oluwole, A.M., Rani, M.R.A., Rohani, J.M., 2015. Commercial bus accident analysis through accident database. J. Transp. Syst. Eng. 2 (1), 7–14.

Quddus, M.A., 2008. Time series count data models: an empirical application to traffic accidents. Accid. Anal. Prev. 40, 1732–1741.

Salifu, N., 2016. Time series analysis of road accidents in Ghana. Dama Int. J. Res. 1, 68–75.

Sanusi, R.A., Adebola, F.B., Adegoke, N.A., 2016. Cases of road traffic accident in Nigeria: a time series approach. Mediter. J. Soc. Sci. 7, 542–552.

Theofilatos, A., Yannis, G., 2014. A review of the effects of traffic and weather characteristics on road safety. Accid. Anal. Prev. 72, 244–256.

Wei, W.W.S., 2006. Time Series Analysis: Univariate and Multivariate Methods. Pearson Education, Boston.

World Health Organization, 2013. Global Status Report on Road Safety 2013. World Health Organization, Geneva.

World Health Organization, 2015. Global Status Report on Road Safety 2015. World Health Organization, Geneva.