International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Time Series Analysis of Road Accidents Using Autoregressive Integrated Moving Average (ARIMA) Model

Joel Cheruiyot Chelule[1], Meshack Kipchumba Ngetich[2], Ayubu Anapapa[3], Herbert Imboga[4]

[1, 2, 4]Jomo Kenyatta University of Agriculture and Technology, Department of Statistics and Actuarial Science, Nairobi, Kenya

[3]University of Eldoret, Department of Mathematics and Computer Science, Eldoret, Kenya

Abstract: ***The road transport industry in Kenya plays a vital role in the life of the majority of her citizens. Many Kenyans utilize different transport modes to reach their various destinations daily. Nearly 3000 people killed on Kenyan roads per year. The objective of this study was a time series analysis of road accidents trend in Kenya using the Autoregressive Integrated Model (ARIMA) model. This study used time series techniques which can better describe and model the accident data. This is achieved using suitable techniques whose performances are subsequently analyzed. The study utilized accident data between the years 2014-2017 obtained from National Transport Safety Authority. In this research project, the time series with Box – Jenkins method applied to 4 years of annual road accident data from 2014 – 2017 to determine the trend of road traffic accident cases and deaths in Kenya. ARIMA models subsequently fitted for accident cases and deaths.***

Keywords: Time Series Analysis, Road Accidents, Autoregressive Integrated Moving Average model (ARIMA)

1. Introduction

The increase in road transport has brought benefits to society in terms of mobility and accessibility; it also, however, has costs. These costs include not only the direct cost of providing transport services such as infrastructure, personnel, equipment costs but also the various indirect costs in terms of the negative impact on the environment such as noise and air pollution, travel delay due to traffic congestion, and the loss of life and property damage as a result of road accidents. This project focused on the significant aspect of road transport activity that is: time series analysis of road accidents and how to fit the model. Road accidents are among the leading cause of death and disability in Kenya, where nearly three thousand people die annually because of accidents related to activities. Despite the efforts to improve road infrastructure, road accidents have continued to occur almost with haste, even with the introduction of the speed control devices, safety belts and other reforms in the sector (Ministry of Transport, 2007).

Road accidents have been acknowledged as one of the adverse element which contributes to the suffocation of the economic growth in the developing countries Kenya being one of them, due to high cost related to them hence causing social and economic concerns. These substantial costs related with road accidents, including human costs (e.g., willingness for someone to pay to avoid pain, grief, and suffering), the direct economic costs of lost output, and the medical costs associated with road accident injuries. Also, costs of damage to vehicles and properties, police costs and administrative costs of accidents insurance, also the combination of factors, including rapid motorization, reduced road and traffic infrastructure, as well as the behavior of road users (Nantulya, 2012). This contrasts with technologically advanced countries where the indices are reducing (Oskam J., 2002). One of the significant challenges faced today is the improvement of the quality of service in the transport sector to make them safe. Despite the enormous economic burden exerted by road traffic accidents, the major causes of accidents in Kenya have not been exhaustively analyzed and modeled to outline the significant causes and their interrelatedness. This emphasizes the need to comprehensively understand the major causes of these accidents and response strategies Kenya can adopt in its endeavor to bring road accidents to a halt of a bare minimum. On May 11, 2011, during the lunch of United Nation Decade of Action for Road Safety 2011-2020, it was discovered that 1.9 million lives would be lost per annum by 2020 worldwide if nothing is done to reduce road accidents. In a keynote address, at the launch, David Cameroon, the U.K Prime Minister stated that "Every six seconds someone is killed or seriously injured on the world's roads." The Russian President, Dimity Medvedev stated "Experts estimate that more than a million people die on the roads each year one in five of whom is a child. More than 50 million people are hurt or seriously injured. The international community, therefore, has a great duty to develop a common strategy and joint action to enhance road safety." This study will analyze the trend of road accidents cases and deaths in Kenya and to fit a suitable ARIMA model for the road accidents data in Kenya. This study uses ARIMA (Box-Jenkins) methodology to develop a time series model for both descriptive and forecasting purposes. This will be achieved using suitable techniques whose performances are subsequently analyzed. The study utilized accident data between the years 2014-2017 obtained from NTSA.

2. Review of Road Accident Models

2.1 Trend of Road Accidents Globally

In 2010, the United Nations General Assembly adopted resolution 64/2551 proclaiming a Decade of Action for Road Safety to stabilize and reduce the increasing trend in road

traffic fatalities. However, a Road accident in high-income countries is expected to fall by 2020, while the converse is true for the developing countries. More than 85% of RTAs occurs in developing countries. The total number of losses realized in developing countries per year exceeded the annual amounts of aid and loans received for development. It has been suggested that the cost to the economy due to RTAs costs an approximately 1- 2% of a country's gross national product (Organization, Global Status Report on Road Safety 2013, 2013)

2.2 Trend of Road Accidents in Kenya

Road accidents are at an alarming rise in Kenya, but very few studies on its causes have been done. However, some thought-provoking facts have been studied and revealed about accidents in Kenya. The figures state the need for quick and detailed research on accidents and their causes to minimize this menace. Kenya, being a developing country, recorded the highest number of accidents. The reasons for many accidents at the marked hotspot include over speeding, careless overtaking and unsafe pedestrian crossing (Daily, 1st August 2018). The number of deaths from road traffic accidents rose from 1,850 in 1990 to 2,830 in 2000 comprising: pedestrians, 40 percent; passengers, 40 percent; drivers, motorcyclists, and bicyclists, 20% (Ministry of Transport, 2007)

2.3 Review of Ordinary Least Square Model

The ordinary least squares method (OLS) is widely applied to estimate regression coefficients of prediction models describing the relationship between road accident and a set of factors describing the underlying transport system, such as many motor vehicles, speed, the volume of the traffic, road design and population size. (Emanalo S., 1977), considered the road transportation system in Zambia and conducted analytical studies to identify the trend of several measures such as the frequency of road accident occurrences and the rate of death resulting from the accident.

2.4 Review of Log-Linear Model

(Kim K., 1995), utilized a log-linear model to explain the role of driver behaviors in the causal sequence that led to a more brutal injury. The study showed that the use of alcohol and lack to use seat belts significantly increased the odds of more severe crashes and injuries. This was also employed when they utilized the Artificial Neural Network (ANN) using Multilayer perception to predict the likelihood of an accident happening at a particular location between the first 40 kilometers along Lagos-Ibadan Express road.

2.5 Time series Models

(Razzaghi, 2013), extended the application of time series analysis to the road safety field and used the data from crashes occurring in Taybad between 2007 and 2011 for investigating the possible patterns of road crashes during the study period, where the time series analysis used a time lag of one month. (Hermans E., 2006), studied the monthly developments in the rate of traffic crashes in Belgium during the period from 1974 to 1999 to identify the trend and investigating the effect of the weather conditions and economy on the road accident crashes rates. (Monfared A., 2013), used Autoregressive Moving Average (ARIMA) models to describe the trend of the death rate of the road accident in Iran, 2004-2011. The analytical studies revealed how powerful the ARIMA technique is in modeling and capturing the variability in a dataset observed at consecutive points of time. (Ofori, 2012), conducted a comparative study between ARIMA and Exponential Smoothing techniques and measured their effectiveness in developing an accurate prediction model for the road crash injuries in Ghana, where the study reported the effectiveness of the ARIMA model over its counterpart the Exponential Smoothing models.

2.6 Review of Box-Jenkins (ARIMA) Models

The ARIMA model is an integration of autoregressive and moving average models, and it is commonly used because of its flexibility. The letter 'I' which lies in the middle of the name 'ARIMA' stands for integration or a differencing operator is needed to make the series stationary (COST329, 2004). The ARIMA model is a potent tool which gives accurate short-range forecasts in time series analyses.

3. Methodology

3.1 Study Design and Population

The study relied on secondary data obtained from the National Transport Safety Authority (NTSA). The data comprises explicitly time series data on daily road traffic accidents covering the period from January 2014 to December 2017. The Authority regularly gathers data on road traffic accident data from hospitals, police stations, forensic medicine, and road organization. To ensure the quality of the collected data, any duplicate or redundant information concerning the road accident was cleaned. Box-Jenkins method used to derive ARIMA models for forecasting the data. This method is preferred because of its high accuracy in forecasting data, especially within a short and medium term period. Also, the model simplicity gives it an advantage of cost and response time, because high cost is required to run and set up complex models (Nihan, 1980).

3.2 Method of Data Analysis

The time-series analysis was applied to model the observed frequency of accident data in the study and to predict future incidences. The Box-Jenkins approach used to develop the best autoregressive integrated moving average (ARIMA) model. The Daily time-series observation is used to increase the prediction power of the model. The ARIMA model will be expressed by ARIMA (p,d,q), where the p,d, and q represents the number of ordinary autoregressive, differences (or integration), and moving average parameters, respectively. In simple term, the p and q are the number of significant lags of the autocorrelation function (ACF) and the partial autocorrelation function (PACF) plots, respectively, and d is the differenced order needed to remove the ordinary non-stationarity in the mean of the error terms.
The Akaike Information Criteria (AIC) was calculated to evaluate the goodness of fit for each model (Box GEP, 2016). This indicator evaluates the model fitness based on

the likelihood model and many parameters. The smaller the size, the better was the model. Finally, the fitted ARIMA model will be used to predict the trend of the people involved in road accident cases and deaths. All the analyses and the forecast will be computed using the R statistical software. The statistical significance will be decided at $p < 0.05$.

3.3Autoregressive Integrated Moving Average (ARIMA)

This is a general model introduced by (Box, 1976) which includes autoregressive as well as moving average parameters, and clearly includes differencing in the formulation of the model. There are three types of parameters in the model are the autoregressive parameters (p), the number of differencing passes (d), and moving average parameters (q). In the notation introduced by Box and Jenkins, models are summarized as ARIMA (p, d, q) e.g

$$\hat{y}_t = \mu + y_{t-1} + \emptyset(y_{t-1} - y_{t-2})$$

3.4 Model Building

3.4.1 Model Identification

Here, the Identification step involves the use of the techniques to determine the values of p, q,and$d$. The values were determined by using the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). In ARIMA (p, d, q) process, the theoretical PACF has non-zero partial autocorrelations at lags 1, 2...p and has zero partial autocorrelations at all lags, while ACF has non zero autocorrelation at lags 1, 2, 3, 4, …,q and zero autocorrelations at all lags.

3.4.2 Estimation Stage

Once a model is identified the next stage ofthe ARIMA model building process is to estimate the parameters ($p, dandq$). When estimating the parameters for the ARIMA (Box- Jenkins) models two approaches were used in the estimation, these are non-linear least squares and maximum likelihood estimation. This study the estimation of the parameters will be done using a statistical package R.

3.4.3 Model Diagnostic Stage

In this stage, different models can be obtained for various combinations of AR and MA individually and collectively, where the best model was obtained with the following diagnostics:

1) Diagnostic of Residuals- here we will use: Time plot of the residuals, the plot of the residuals ACF, the normal Q-Q plot, and testing the model for adequacy
2) Tests of Significant of coefficients

Before performing the time series analysis, the quality of the collected data will be assessed in terms of data integrity. Based on data integrity, the daily and monthly period will be chosen. It is necessary for a stationary mean and variance to be established. In order to remove the seasonality variation and trend from the observed time series, seasonal differencing and order differencing, respectively, will be applied to the data. The patterns of the autocorrelation function (ACF) and the partial autocorrelation function (PACF) plots will be used to identify best models. In this study, several different models will be identified through analysis of the ACF and PACF plots, including AR (autoregressive), MA (moving average), ARMA (autoregressive moving average), ARIMA (autoregressive moving integrated moving average), and SARIMA (seasonality autoregressive moving integrated moving average).

4. Results and Analysis

4.1 Data Presentation

Accident data on the Kenyan highways for the period 2014-2017 were compiled from the NTSA, which involves a number of accident cases, the number of people involved and the number of people died.

4.2 Descriptive analysis of Road Accident Cases

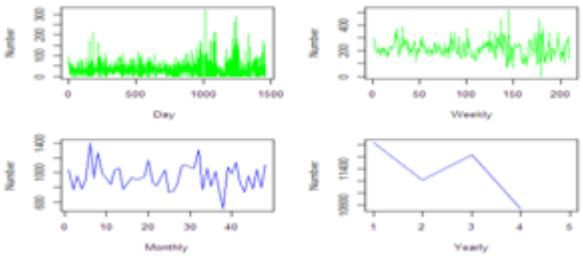

Figure 4.2: Daily, Weekly, Monthly and Yearly Time Plot for the Number of People Involved in Road Accidents in Kenya for the year 2014-2017.

The figure above shows a time plot for total people involved in road accidents in Kenya from 2014-2017. An irregular pattern is observed on the daily data that is on the top left, the same also can be observed on the top right that shows a weekly number of people involved in road accidents. The bottom left displays monthly accident data; it can be seen that there is a systematic pattern on monthly data this can be attributed to the seasonal effects. The bottom right figure shows the yearly number of people involved in road accidents. It can be seen that in the year 2014, the number of people involved in RTA was very high compares to the succeeding years. There was a drastic fall in 2015, for the number of people involved in RTA. Then it steadily rises in 2016 before falling in 2017.

4.3 Descriptive analysis of Road Accident deaths

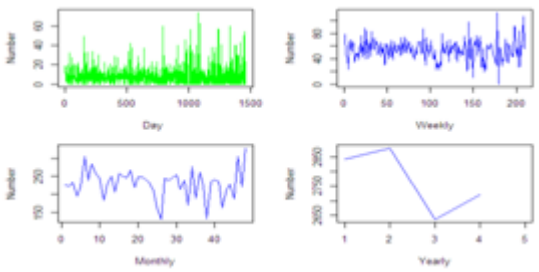

Figure 4.3: Daily, Weekly, Monthly and Yearly Time Plot for the Number of People Died in Road Accidents in Kenya for the year 2014-2017

The above figure shows the number of fatalities in road accidents in Kenya for the period 0f 2014-2017. An irregular pattern is observed on the daily data that is on the top left, the same also can be observed on the top right that shows the

weekly number of people died in road accidents. The bottom left displays monthly accident data; it can be seen that there is a systematic pattern on monthly data this can be attributed to the seasonal effects. The bottom right figure shows a yearly number of people involved in road accidents. It can be seen that in the year 2014, the number of people who died in RTA was very. Then there was a rise in 2015, for the number of people who dies in RTA. Then it sharply falls in 2016 before it started to rise again in 2017. This shows that there is a rising trend in the number of people dying as a result of road accidents in Kenya.

Testing the hypothesis

We will try to establish if the mean number of accidents and the number of people who died in a road accident are the same for Day of the week, quarterly and monthly data

ANOVA for Day of the Week Accidents Data

The mean for the number of people involved in RTA for the day of the week is given below from the data

Fri	Mon	Sat	Sun	Thu	Tue	Wed
29.92344	34.70192	33.04785	34.84211	27.32057	30.96635	26.76555

H_0: No difference in mthe ean number of people involved in the day of the week

H_1: There is a significant difference in mthe ean number of people involved in accidents during the day of the week

Table 4.1 ANOVA Table for Total Number of People Involved During the Day of the Week

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
data$ Day.of.the.week	6	13623	2270	2.264	0.0352	*
Residuals	1454	1457927	1003			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Inference and Conclusion

Since p-value <0.05, we, therefore, reject the null hypothesis. There is a statistically significant difference in the mean number of people involved in RTA during the day of the week. We, therefore, compute Tukey HSD

Table 4.2: TukeyHSD Table for Total Number of People Involved During the Day of the Week

Tukey multiple comparisons of means				
95% family-wise confidence level				
	diff	lwr	upr	p adj
Mon-Fri	4.778478	-4.37811	13.93507	0.720014
Sat-Fri	3.124402	-6.0212	12.27001	0.952138
Sun-Fri	4.91866	-4.22694	14.06426	0.690218
Thu-Fri	-2.60287	-11.7485	6.542732	0.980669
Tue-Fri	1.042901	-8.11369	10.19949	0.999885
Wed-Fri	-3.15789	-12.3035	5.987708	0.949643
Sat-Mon	-1.65408	-10.8107	7.502513	0.998356
Sun-Mon	0.140182	-9.01641	9.296771	1
Thu-Mon	-7.38135	-16.5379	1.77524	0.20772
Tue-Mon	-3.73558	-12.9031	5.431984	0.893141
Wed-Mon	-7.93637	-17.093	1.220216	0.139568
Sun-Sat	1.794258	-7.35135	10.93986	0.997388
Thu-Sat	-5.72727	-14.8729	3.41833	0.514975
Tue-Sat	-2.0815	-11.2381	7.075088	0.994118
Wed-Sat	-6.2823	-15.4279	2.863306	0.397185
Thu-Sun	-7.52153	-16.6671	1.624072	0.187484
Tue-Sun	-3.87576	-13.0323	5.28083	0.874364
Wed-Sun	-8.07656	-17.2222	1.069048	0.124408
Tue-Thu	3.645772	-5.51082	12.80236	0.90337
Wed-Thu	-0.55502	-9.70063	8.590579	0.999997
Wed-Tue	-4.2008	-13.3574	4.955793	0.825734

This output indicates that the differences Monday-Friday, Sat-Friday, all through to Wednesday-Tuesday are significant. A more "easy" way to interpret this output is visualizing the confidence intervals for the mean differences. The mean for the number of people died in RTA for the day of the week is given below from the data

Fri	Mon	Sat	Sun	Thu	Tue	Wed
7.215311	8.317308	8.311005	9.215311	6.444976	7.048077	6.531100

H_0: No difference in mthe ean number of people died in the day of the week

H_1: There is a significant difference in the mean number of died in accidents during the day of the week

Table 4.3: ANOVA Table for Total Number of Deaths during the Day of the Week

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Data $ Day of the week	6	1369	228.2	3.342	0.00283**
Residuals	1454	99302	68.3		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Inference and Conclusion

Since p-value <0.05, we,therefore, reject the null hypothesis. There is a statistically significant difference in the mean number of people died in RTA during the day of the week. We, therefore, calculate Tukey HSD

Table 4.4: TukeyHSD Table for Total Number of People Involved During the Day of the Week

Tukey multiple comparisons of means				
95% family-wise confidence level				
	diff	lwr	upr	p adj
Mon-Fri	1.101996688	-1.2877066	3.4917000	0.8221821
Sat-Fri	1.095693780	-1.2911425	3.4825300	0.8253106
Sun-Fri	2.000000000	-0.3868363	4.3868363	0.1695456
Thu-Fri	-0.770334928	-3.1571712	1.6165013	0.9636533
Tue-Fri	-0.167234082	-2.5569374	2.2224692	0.9999935
Wed-Fri	-0.684210526	-3.0710468	1.7026257	0.9799417
Sat-Mon	-0.006302908	-2.3960062	2.3834004	1.0000000
Sun-Mon	0.898003312	-1.4917000	3.2877066	0.9253617
Thu-Mon	-1.872331616	-4.2620349	0.5173717	0.2381834
Tue-Mon	-1.269230769	-3.6617977	1.1233362	0.7040233
Wed-Mon	-1.786207214	-4.1759105	0.6034961	0.2922987
Sun-Sat	0.904306220	-1.4825300	3.2911425	0.9225373
Thu-Sat	-1.866028708	-4.2528650	0.5208075	0.2405825
Tue-Sat	-1.262927862	-3.6526312	1.1267755	0.7077602
Wed-Sat	-1.779904306	-4.1667406	0.6069320	0.2950841
Thu-Sun	-2.770334928	-5.1571712	-0.3834987	0.0111874
Tue-Sun	-2.167234082	-4.5569374	0.2224692	0.1046639
Wed-Sun	-2.684210526	-5.0710468	-0.2973743	0.0160309
Tue-Thu	0.603100847	-1.7866025	2.9928042	0.9896825
Wed-Thu	0.086124402	-2.3007119	2.4729607	0.9999999
Wed-Tue	-0.516976445	-2.9066798	1.8727269	0.9955127

This output indicates that the differences Monday-Friday, Sat-Friday, all through to Wednesday-Tuesday are significant. A more "easy" way to interpret this output is visualizing the confidence intervals for the mean differences.

ANOVA for Monthly Accidents Data

The monthly mean for the total number of people involved in RTA is given below:

April	August	December	February	January	July
31.56667	38.07258	33.84677	24.78761	28.84677	29.62903

June	March	May	November	October	September
35.43333	30.12903	32.65323	28.05000	31.11290	28.22500

H_0: No difference in the mean number of people involved in accidents monthly

H_1: There is a significant difference in the mean number of people involved in accidents monthly

Table 4.5: ANOVA Table for Monthly Total Number of People Involved in RTA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
data$ Month of the Year	11	17168	1561	1.555	0.106
Residuals	1449	1454382	1004		

Inference and Conclusion

Since p-value >0.05, we, therefore, fail to reject the null hypothesis. There is no statistically significant difference in the mean number of people involved in RTA monthly. The monthly mean for the total number of people died in RTA is given below:

April	August	December	February	January
7.725000	8.072581	8.306452	6.115044	6.951613

July	June	March
7.161290	8.266667	7.838710

May	November	October	September
7.677419	6.900000	8.653226	7.191667

H_0: No difference in the mean number of deaths monthly

H_1: There is a significant difference in the mean number of deaths monthly

Table 4.6: ANOVA Table for Monthly Total Number of People Deaths in RTA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Data $ Month of the Year	11	694	63.06	0.914	0.526
Residuals	1449	99977	69.00		

Inference and Conclusion

Since p-value >0.05, we, therefore, fail to reject the null hypothesis. There is no statistically significant difference in the mean number of deaths monthly.

ANOVA for Quarterly Accidents Data

The Quarterly mean for the total number of people involved in RTA is given below:

Q1	Q2	Q3	Q4
28.01662	33.21154	32.01630	31.03533

H_0: No difference in the mean number of people involved in accidents quarterly

H_1: There is a significant difference in the mean number of people involved in accidents quarterly

Table 4.7: ANOVA Table for Quarterly Total Number of People Involved in RTA

	f	Sum Sq	Mean Sq	F value	Pr(>F)
data $ Quarter	3	5365	1788	1.777	0.15
Residuals	1457	1466185	1006		

Inference and Conclusion

Since p-value >0.05, we, therefore, fail to reject the null hypothesis. There is no statistically significant difference in the mean number of people involved in RTA quarterly. The quarterly mean for the total number of people died in RTA is given below:

Q1	Q2	Q3	Q4
6.994460	7.887363	7.478261	7.964674

H_0: No difference in the mean number of deaths quarterly

H_1: There is a significant difference in the mean number of deaths quarterly

Table 4.8: ANOVA Table for Monthly Total Number of People Deaths in RTA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
data$ Quarter	3	216	72.14	1.046	0.371
Residuals	1457	100455	68.95		

Inference and Conclusion

Since p-value >0.05, we, therefore, fail to reject the null hypothesis. There is no statistically significant difference in the mean number of deaths quarterly.

4.4 Trend Differencing For the number of People Involved in RTA

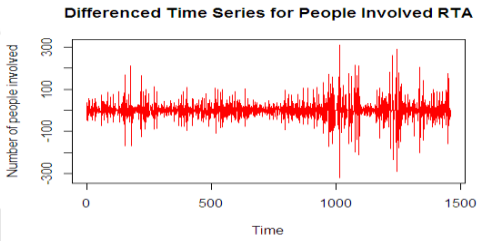

Figure 4.4: First difference of the Number of People Involved in Road Accident.

A transformation of the Road accident cases data using the first differencing method is performed to remove the trend component in the original accident data cases which are shown in Figure 4.3. The observations move irregularly but revert to its mean value and the variability is also approximately constant. The total number of people involved in RTA data now looks to be approximately stable. The following are the ACF and PACF of the total number of people involved in RTA

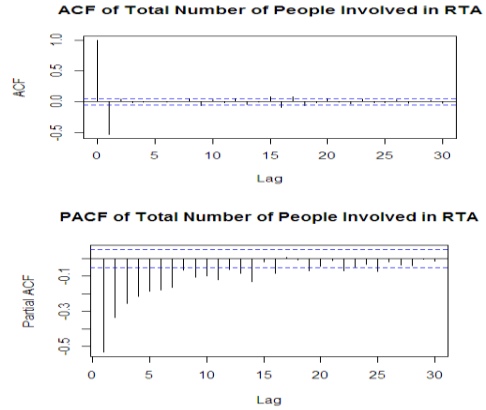

Figure 4.5: ACF and PACF plots of the first differencing of the accident data cases

The top part of Figure 4.5 shows the autocorrelation function of the first differencing of the motorway accident data at various lags and the bottom part is the partialautocorrelation function of the first differencing of the road accident data also at different lags.

Comparing the autocorrelations with their error limits, the only significant autocorrelation is at lag 2, indicating an MA (2) behavior. Similarly, it's geometric at partial autocorrelations are significant, indicating an AR (1) but applying the principle of parsimony we use AR (0). The following models are suggested;

- ARIMA (0,1,2)
- ARIMA (1,1,2)
- ARIMA (2,1,2)

To select the best model for forecasting into the future, each model is assessed based on its parameter estimates, the corresponding diagnostics of the residuals and the AIC.

4.5 Model Selection for the Data

4.5.1 Parameter Estimates ARIMA Models

ARIMA (0,1,2)
Call:
arima(x = diffdata1, order = c(0, 1, 2))

Coefficients:

	ma1	ma2
	-1.9744	0.9747
s.e.	0.0049	0.0048

sigma^2 estimated as 1030: log likelihood = -7137.73, aic = 14281.46

Tests for coefficients

z test of coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
ma1	-1.9743872	0.0048780	-404.75	<2.2e-16 ***
ma2	0.9747217	0.0047511	205.16	<2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

ARIMA (1,1,2)
Call:
arima(x = diffdata1, order = c(1, 1, 2))

Coefficients:

	ar1	ma1	ma2
	-0.0796	-1.9910	0.9910
s.e.	0.0265	0.0034	0.0031

sigma^2 estimated as 1007: log likelihood = -7125.06, aic = 14258.12

Tests for coefficients

z test of coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
ar1	-0.0796037	0.0265375	-2.9997	0.002703 **
ma1	-1.9910042	0.0033525	-593.8849	< 2.2e-16 ***
ma2	0.9910087	0.0031380	315.8088	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

ARIMA (2,1,2)

Call:
arima(x = diffdata1, order = c(2, 1, 2))

Coefficients:

	ar1	ar2	ma1	ma2
	-0.0899	-0.0303	-1.9732	0.9733
s.e.	0.0266	0.0266	0.0058	0.0057

sigma^2 estimated as 1016: log likelihood = -7128.97, aic = 14267.93

Tests for coefficients

z test of coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
ar1	-0.0898906	0.0266007	-3.3793	0.0007268
ar2	-0.0303363	0.0265895	-1.1409	0.2539052
ma1	-1.9731981	0.0058459	-337.5339	< 2.2e-16
ma2	0.9732671	0.0057204	170.1393	< 2.2e-16

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

4.5.2 Diagnostic for ARIMA Models

Test for the significance for ARIMA (0,1,2)

Box-Ljung test
data: arimaModel_1$residuals
X-squared = 46.554, df = 20, p-value = 0.0006762
ma1 ma2
0 0

z test of coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
ma1	-1.9743872	0.0048780	-404.75	< 2.2e-16 ***
ma2	0.9747217	0.0047511	205.16	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The p – values for the Ljung-Box statistics is not significant at any positivelag. That is all p – values are less than 0.05.

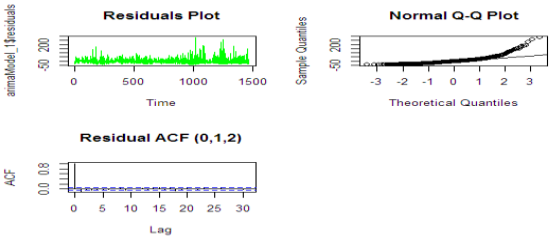

Figure 4.6: Diagnostic plot of ARIMA (0,1,2)

Diagnostics of the residuals from ARIMA (0, 1, 2) is shown in Figure 4.7 above.

a) The residuals plot shows no obvious pattern and looks like an i.i.d. of mean zero with few outliers.
b) The ACF of the residuals plot shows no significant residual autocorrelation for the ARIMA (0, 1, 2) model.
c) The normal Q-Q plot of the residuals doesn't look too bad, so the assumption of normally distributed residuals look okay.

Test for the significant for ARIMA (1,1,2)

ar1	ma1	ma2
0.002702688	0.000000000	0.000000000

z test of coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
ar1	-0.0796037	0.0265375	-2.9997	0.002703 **
ma1	-1.9910042	0.0033525	-593.8849	< 2.2e-16 ***
ma2	0.9910087	0.0031380	315.8088	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Box-Ljung test

data: arimaModel_2$residuals

X-squared = 34.275, df = 20, p-value = 0.02432

The p – values for the Ljung-Box statistics is not significant at any positivelag. That is all p – values are less than 0.05.

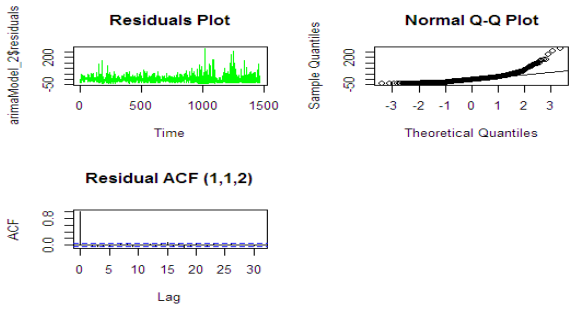

Figure 4.7: Diagnostic plot of ARIMA (1,1,2)

Diagnostics of the residuals from ARIMA (1, 1, 2) is shown in Figure 4.7 above.

a) The residuals plot shows no obvious pattern and looks like an i.i.d. of mean zero with few outliers.
b) The ACF of the residuals plot shows no significant residual autocorrelation for the ARIMA (1, 1, 2) model.
c) The normal Q-Q plot of the residuals doesn't look too bad, so the assumption of normally distributed residuals look okay.

Test for the significant for ARIMA (2,1,2)

z test of coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
ar1	-0.0898906	0.0266007	-3.3793	0.0007268 ***
ar2	-0.0303363	0.0265895	-1.1409	0.2539052
ma1	-1.9731981	0.0058459	-337.5339	< 2.2e-16 ***
ma2	0.9732671	0.0057204	170.1393	<2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Box-Ljung test

data: arimaModel_3$residuals

X-squared = 36.989, df = 20, p-value = 0.01174

The p – values for the Ljung-Box statistics is not significant at any positive lag. That is all p – values are less than 0.05.

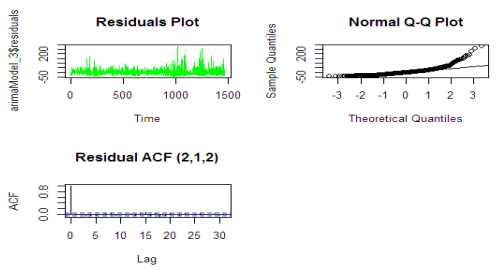

Figure 4.8: Diagnostic plot of ARIMA (2,1,2)

Diagnostics of the residuals from ARIMA (1, 1, 2) is shown in Figure 4.7 above.

a) The residuals plot shows no obvious pattern and looks like an i.i.d. of mean zero with few outliers.
b) The ACF of the residuals plot shows no significant residual autocorrelation for the ARIMA (1, 1, 2) model.
c) The normal Q-Q plot of the residuals doesn't look too bad, so the assumption of normally distributed residuals look okay.

4.5.3 Selection of the Best Model for Forecasting Number of People Involved in RTA

Table 4.9: Summary of the ARIMA Models Estimates and Standard Error

Model	Testing on Parameter Estimates			
	Parameter	Estimates	S.E	Significant
ARIMA (0,1,2)	constant	-1.9744	0.0049	Not significant
	Ma1	-1.9744	0.0049	Not significant
	Ma2	0.9747	0.0048	Not significant
ARIMA (1,1,2)	Ar1	-0.0796	0.0265	Not significant
	Ma1	-1.9910	0.0034	Not significant
	Ma2	0.9910	0.0031	Not significant
ARIMA (2,1,2)	Ar1	-0.0899	0.0266	Not significant
	Ar2	-0.0303	0.0266	Significant
	Ma1	-1.9732	0.0058	Not significant
	Ma2	0.9910	0.0057	Not significant
DIAGNOSTICS				
	ARIMA (0,1,2)	ARIMA (1,1,2)	ARIMA (2,1,2)	
AIC	14281.46	14267.93	14258.12	

From the above table, we can see that the AIC for ARIMA(0,1,2) is 14281.46, for ARIMA(1,1,2) is 14258.12 and for ARIMA(2,1,2) is 14268.93. The AICis good for all the models but they favor ARIMA (1, 1, 2),model. For the three ARIMA models, the residuals plot shows no obvious pattern and looks like an i.i.d. of mean zero with few outliers, The ACF of residuals plot shows no significant residual autocorrelation for the, and The normal Q-Q plot of the residuals doesn't look too bad, so the assumption of normally distributed residuals look okay. Hence the normality assumption is satisfied.

From the discussion above it is clear that ARIMA (1,1,2) model is the best model for forecasting the motorway accident mortality.

4.7 Fitting the ARIMA Model for the Total Number of People Involved in RTA

ARIMA (1,1,2) is the best model for forecasting for the total number of people involved in road accidents cased in Kenya. The model, therefore, is given as:

$$y_t = y_{t-1} + \emptyset(y_{t-1} - y_{t-2}) - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2}$$

In terms of observed series we have;

$$y_t - y_{t-1} = \emptyset(y_{t-1} - y_{t-2}) - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2}$$

The point estimate of each parameter of ARIMA (1, 1, 2) from table 4.1 are as follows:

$$\emptyset = -0.796, \theta_1 = -1.9910, \theta_2 = 0.9910$$

The fitted ARIMA (1,1,2) model is therefore given as:

$$y_t = y_{t-1} - 0.796(y_{t-1} - y_{t-2}) - 1.9910\varepsilon_{t-1} + 0.9910\varepsilon_{t-2}$$

Where$\varepsilon_t$has an estimated variance of 1007.

4.5.1 Selection of the Best Model for Forecasting Death Cases

Table 4.10: Summary of the ARIMA Models Estimates and Standard Error

Model	Testing on Parameter Estimates			
	Parameter	Estimates	S.E	Sig
ARIMA(1,1,0)	MA1	-0.6876	0.0190	Not significant
	constant	-1.9744	0.0049	Not significant
	Ma1	-1.0000	0.0019	Not significant
ARIMA(1,1,1)	Ar1	-0.5280	0.0222	Not significant
	Ma1	-1.0000	0.0019	Not significant
DIAGNOSTICS				
	ARIMA (0,1,2)	ARIMA (1,1,2)	ARIMA (2,1,2)	
AIC	12148.46	11460.62	10985.60	

From the above results we can see that the AIC for ARIMA(1,1,0) is 12148.46, for ARIMA(0,1,1) is 11460.62 and for ARIMA(1,1,1) is 10985.60. The plots for residuals ACF of all models show significantly. The normal Q-Q plot of the residuals indicates that residuals are located on the straight line except a few that are deviating from the normality. Hence the normality assumption is satisfied and appeared to be normally distributed.

The AIC is good for all the models but they favor ARIMA (1, 1, 1),model. From the discussion above it is clear that ARIMA (1,1,1) model is the best model for forecasting the road accident death.

4.8 Fitting the Model for the Number of Deaths

ARIMA (1,1,1) model is the best model for forecasting the motorway accident death, this is a non- seasonal model with one AR term and one MA term. The model in terms of the differenced series $\boldsymbol{xt}$is given by:

$$y_t = \emptyset y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

In terms of observed series, the model is given as

$$y_t - y_{t-1} = \emptyset(y_{t-1} - y_{t-2}) - \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

The fitted ARIMA (1,1,1) model is therefore given as:

$$y_t - y_{t-1} = -0.5280(y_{t-1} - y_{t-2}) - 1.00\varepsilon_{t-1} + \varepsilon_t$$

Where$\varepsilon_t$has an estimated variance of 108.

Conclusion and Recommendations

A road traffic accident in Kenya is increasing at an alarming rate and has raised majorconcerns. The NTSA recognizes the contributions road safety researches makes tothe development of accident reduction initiatives. It is against this background that thisthe research was carried out to analyze the trend of RTA cases anddeaths and to develop a time series ARIMA models to predict two years of road accident cases and deaths along our roads.Time series analysis of the data from the years 2014-2017 showed that patterns ofroad accident cases that is the total number of people involved and deaths are increasing along tour roads. ARIMA models were subsequently developed for the total number of people involved in road accidents and the number of deaths throughout 2014-2017, after identifying various tentative models. ARIMA (1,1, 2) was identified to be a suitable model for forecasting into the future of the number of people involved in the accident cases while ARIMA(1,1,1) was found to be a suitable model for the accident deaths cases in Kenya.

The ARIMA (1,1,2) model was recommended for forecasting the total number of people involved in road accidents while ARIMA(1,1,1) was recommended for road accidents death forecasting along the Kenyan roads, but the following precautionary measures should be taken into consideration to prevent the increasing forecast values of these models: Enforcement of traffic safety campaigns, proper maintenance of the roads, strict adherence to road traffic rules. The models should not be used to forecast a long time ahead (preferably a maximum of 9 years). This is because long periods could lead to arbitrary large forecasts values. Finally, it is also recommended that further study should be done to look for more appropriate models that can take care of drastic government interventions.

References

[1] Authority, N. T. (2018). *National Transport Safety Authority*. Retrieved from NTSA web: www.ntsa.go.ke
[2] Box GEP, J. G. (2016). Time series analysis. In *forecasting and control* (pp. 179–209). New Jersey: Academic.
[3] Box, G. E. (1976). *Time series analysis forecasting and control (Revised Ed.).* Oakland, California: Holden-Day, Inc.
[4] COST329. (2004). Models for traffic and safety development. In the *Final Report of The Action.* Luxembourg: European Commission.
[5] Daily, N. (1st August 2018). *How to stay safe on Kenya's deadliest roads.* Nairobi: Nation Media Group.
[6] Emanalo S. (1977). *Road Traffic Survey In Lusaka.* Lusaka.
[7] Hermans E., W. G. (2006). Frequency and Severity of Belgian Road Traffic Accidents Accidents Studied by State-Space Methods. In *Bureau of Transportation Statistics Vol 9* (pp. 63-76).
[8] Kim K., N. L. (1995). Accident Analysis and Prevention.
[9] Ministry of Transport, K. (2007). National Road Safety Action Plan.
[10] Monfared A., S. B. (2013). Prediction of fatal road traffic crashes in Iran using the Box-Jenkins time series model. *Journal of Asian Scientific Research Vol 3 no 4*, 425-430.
[11] Montella A., C. L. (2008). Crash Prediction Models for Rural Motorways. *Transportation Research Board ISSN: 0361-1981*.
[12] Nantulya, V. &. (2012). The neglected epidemic. *Br. Med. Journal, 324*, pp. 1139-1141.
[13] Ofori, T. A. (2012). Statistical Models for Forecasting Road Accident Injuries in Ghana. *International Journal of Research in Environmental Science and Technology, vol. 2, no. 4*, 143-149.
[14] Organization, W. H. (2013). Global Status Report on Road Safety 2013.

[15] Oskam J., K. J. (2002). The Groningen trauma study. Injury patterns in a Dutch trauma centre. *Eur. J. Emergency Med., 1*, pp. 167-172.

[16] Razzaghi, B. A. (2013). Assessment of Trend and Seasonality in Road Accident Data: An Iranian Case-Study. In *Health Policy and Management Vol 1* (pp. 51-55).