European Journal of Academic Essays 2(3): 13-31, 2015
ISSN (online): 2183-1904
ISSN (print): 2183-3818
www.euroessays.org

The Comparison of the Performance **o**f ARIMA **a**nd MA Model Selection **o**n Road Accident Data **i**n Nigeria

<Balogun, Oluwafemi Samson>[*1], <Oguntunde, Pelumi Emmanuel>[2], <Akinrefon, Adesupo Adeoye>[1],

< Modibbo, Umar Mohammed>[1]

[1]Department of Statistics and Operations Research, Modibbo Adama University of Technology, P.M.B. 2076, Yola,

Adamawa State, Nigeria.

stapsalms@yahoo.com

[2] Department of Mathematics, Covenant University, Ota, Ogun State. Nigeria.

peluemman@yahoo.com

Abstract: In this research work, time series model selection was performed by given consideration for a number of models that most suitable for the incidence of accident cases in Nigeria. Among the candidate models considered are the Autoregressive Integrated Moving Average (ARIMA) and Moving Average (MA) models each at various parameters specifications. Results from this work showed that the best models that are suitable to describe the accident cases in Nigeria are the ARIMA(3,1,1) and MA(0,1,2) according to the Mean Square Error (MSE) and Akaike Information Criteria (AIC). National data set on cases of accident in Nigeria primarily collected by Federal Road Safety Commission (FRSC), Nigeria from 2004 to 2011 was employed in this research.
Keywords: Autocorrelation, Autovariance, Box-Jenkins, ARIMA, MA and AIC.

1. Introduction

Having being disturbed by the unpleasant trend in the nation's road traffic system, the Federal Government of Nigeria initiated and established the Federal Road Safety Commission (FRSC) to check the alarming increase in the number of road traffic in Nigeria. The FRSC was established by Decree 45 of 1988, as the lead agency in Nigeria on road safety administration and management. The vision of the commission is to eradicate road traffic crashes and create safe motoring environment in Nigeria. Missions includes regulate, enforce and coordinate all road traffic and safety management activities through:

Sustained public enlightenment, effective patrol operations, prompt rescue services, improved vehicle administration, robust data management and promotion of stakeholder cooperation.

Within the provision of its enabling Act, the functions of the FRSC are as follows:

(a) Preventing and minimizing road traffic accidents.
(b) Clearing obstructions on the highways.
(c) Educating drivers, motorists and other members of the public on the proper use of the highways.
(d) Providing prompt attention and care to victims of road traffic accidents.
(e) Conducting researches into causes of road traffic accidents.
(f) Determining and enforcing speed limits for all categories of roads and vehicles.
(g) Co-operating with bodies, agencies and group engaged in road safety activities or the prevention of highway accident.

In particular the commission is charged with the responsibilities for:

***Corresponding Author:** Balogun Oluwafemi Samson, Department of Statistics and Operations Research, Modibbo Adama University of Technology, Yola, Adamawa State, *stapsalms@yahoo.com*, +2348035217339

1. Preventing or minimizing accidents on the highway.
2. Clearing obstructions on any part of the highways.
3. Educating drivers, motorists and other members of the public generally on the proper use of the highways.
4. Designing and producing the driver's license to be used by various categories of vehicle operators.
5. Determining, from time to time, the requirements to be satisfied by an applicant for a driver's license.
6. Designing and producing vehicle number plates.
7. The standardization of highway traffic codes.
8. Giving prompt attention and care to victims of accidents.
9. Conducting researches into causes of motor accidents and methods of preventing them and putting into use the result of such researches.
10. Determining and enforcing speed limits for all categories of roads and vehicles and controlling the use of speed limiting devices
11. Cooperating with bodies or agencies or groups in road safety activities or in prevention of accidents on the highways.
12. Making regulations in pursuance of any of the functions assigned to the Corps by or under this Act.
13. Regulating the use of sirens, flashers and beacon lights on vehicles other than ambulances and vehicles belonging to the Armed Forces, Nigeria Police, Fire Service and other Para-military agencies.
14. Providing roadside and mobile clinics for the treatment of accident victims free of charge.
15. Regulating the use of mobile phones by motorists.
16. Regulating the use of seat belts and other safety devices;
17. Regulating the use of motorcycles on the highway;
18. Maintaining the validity period for drivers' licenses which shall be three years subject to renewal at the expiration of the validity period; and

In exercise of the functions, members of the Commission shall have power to arrest and prosecute persons reasonably suspected of having committed any traffic offence.

The aims of this research is to examine the pattern of road accident in Nigeria, to estimate different models of autocorrelation, using the various models (ARIMA and MA models) and to compare the models to obtain the model that best fit.

The data used in this research work is a secondary data collected from the National Head quarter Federal Road Safety Commission office (FRSC) of Nigeria. The data covers the monthly road accidents in Nigeria for a period of 8 years (2004-2011). This research work is concerned with the comparison of some models take into account of autocorrelation. The models compared are the Moving-average models MA(q) and the Auto-regressive integrated moving average model ARIMA(p,d,q).

Time series was originated in 1807 by French Mathematician name FOURIER, who claimed that any Series could be approximated as the sum of the Sine and Cosine terms. In 1960 Schituster used Fourier's idea to estimate the length periodicities and utilized peridogram analysis in his research.

According to [1] he defines time series as a set of observations taken at a specified time usually at equal interval.

According to [2] he define time series as a statistical series which tell us how data has been behaving in the past.

According to [3], he defines time series as a collection of observation segmental in time at regular intervals.

The usage of time series models is in twofold:

(1) To obtain an understanding of the underlying forces and structure that produced the observed data, and

(2) To fit a model and proceed to forecasting, monitoring or even feedback and feed forward control.

Time Series Analysis's includes: Economic Forecasting, Sales Forecasting, Budgetary Analysis, Stock Market Analysis, Yield Projections, Process and Quality Control, Inventory Studies, Workload Projections, Utility Studies, Census Analysis, and many, many more...

1.1 Types of Time Series

There are 3 types of time series which are:

(1) **Continuous Time Series:** This involves Hydrological parameters which are often continuously recorded. This occurs either on the record sheet of a chart recorder, or a data logger is used. A data logger typically records the data either at fixed time intervals or after a certain change in the Y-value has taken place. Despite this sampling, the data are interpreted as if they were continuous data. The data are recorded so that the information content due to the continuity is retained. (E.g. a precipitation event or precipitation free).

(2) Interval Time Series: An interval time series does not contain values for points in time but rather for particular intervals of time. These time intervals can be equidistantly or randomly distributed in time. Equidistant in terms of years or months still means that the actual intervals have different lengths. A typical equidistant time series is a daily total series, where each value is for an interval of 24 hours.

(3) Momentary Time Series: The momentary time series is the rarest form of time series. In contrast to the other time series, a momentary time series is only defined for a discrete set of points in time. The time series does not contain any information for the time between these points. Interpolation is not meaningful, and the value function thus has the value undefined for these points. An example of a momentary time series is the series of local maxima of a precipitation time series. The set of points in time is made up of randomly distributed points in time. There is no information for all other points in time.

1.2 Time Series Models

A time series consists of observations at discrete equi-spaced intervals of time. For example, Accidents in month "t" could be denoted as X_t and in the previous month by X_{t-1}. Typically, the objective of time series analysis is to forecast future values of X (such as X_{t+1}) based on present and past values of X and perhaps also on explanatory variables such as accidents

A model in which future values are forecast purely on the basis of past values of the time series is called an Autoregressive (AR) process.

A model in which future values are forecast purely on the basis of past shocks (or noise or random disturbances) is called a Moving average (MA) process.

A model that uses both past values of the time series and past shocks is called an autoregressive-moving average (ARMA) process.

These models assume that the time series is stationary - that is the series fluctuates around a time invariant mean, and the variance and autocovariance i.e. covariance between X_t and X_{t-s} (for all values of s) do not vary with time. In practice, most time series need to be transformed to achieve stationarity. To stabilize variance a logarithm transform is often used - appropriate where the variance of the series increases in proportion to the mean. To stabilize the mean, differencing is usually employed. For example, first order differencing is $Z_t = X_t - X_{t-1}$. First order differencing eliminates "drift" but it often needs to be applied twice to eliminate trend. Seasonal differencing is often necessary too. An ARMA model of a differenced series is called an ARIMA model, where the 'I' stand's for Integrated because the output needs to be anti-differenced or integrated, to forecast the original series.

Collectively, these models along with the process of identification, fitting, and diagnostic checking are called Box Jenkins models [4].

One fundamental goal of statistical modeling is to use the simplest model possible that still explains the data. This is known as principle of parsimony [5].

1.2.1 ARIMA

Early attempts to study time series, particularly in the 19th century, were generally characterized by the Idea of a deterministic world. It was the major Contribution of [6] which launched the notion of stochasticity in time series by postulating that every time series can be regarded as the realization of a stochastic process. Based on this simple idea, a number of time series methods have been developed since then.

Workers such as Slutsky, Walker, Yaglom, and Yule first formulated the concept of autoregressive (AR) and moving average (MA) models. Wold's decomposition theorem led to the formulation and solution of the linear forecasting problem of [7].

Since then, a considerable body of literature has appeared in the area of time series, dealing with parameter estimation, identification, model checking, and forecasting; e.g., [8] for an early survey. The publication Time Series Analysis: Forecasting and Control by [9] integrated the existing knowledge.

1.2.2 AUTOREGRESSIVE

Autoregressive (AR) models were first introduced by [10].They were consequently supplemented by [11] presented Moving Average (MA) schemes. It was [12], however, who combined both AR and MA schemes and showed that ARMA processes can be used to model all stationary time series as long as the appropriate order of p, the number of AR terms, and q, the number of MA terms, was appropriately specified. This means that any series x t can be modeled as a combination of past x(t) values and/or past e(t) errors.

The utilization of the theoretical results suggested by Wold, to model real life series did not become possible until the mid 1960s when computers, capable of performing the required calculations became available and economical. [13] original edition [9] popularized the use of ARMA models through the following:

(a) Providing guidelines for making the series stationary in both its mean and variance,

(b)Suggesting the use of autocorrelations and partial autocorrelation coefficients for determining appropriate values of p and q (and their seasonal equivalent P and Q when the series exhibited seasonality),

(c) providing a set of computer programs to help users identify appropriate values for p and q, as well as P and Q, and estimate the parameters involved and

(d) once the parameters of the model were estimated, a diagnostic check was proposed to determine whether or not the residuals e(t) were white noise, in which case the order of the model was considered final (otherwise another model was entertained in (b) and steps (c) and (d) were repeated). If the diagnostic check showed random residuals then the model developed was used for forecasting or control purposes assuming of course constancy that is that the order of the model and its non-stationary behavior, if any, would remain the same during the forecasting, or control, phase.

The approach proposed by Box and Jenkins came to be known as the Box-Jenkins methodology to ARIMA models, where the letter "I", between AR and MA, stood for the word "Integrated". ARIMA models and the Box-Jenkins methodology became highly popular in the 1970s among academics, in particular when it was shown through empirical studies ([14]; [15]; [16]; [17]; [18], for a survey see [19]) that they could outperform the large and complex econometric models, popular at that time, in a variety of situations.

2. Methodology

If future values can be predicted exactly from past values, then a series is said to be deterministic. One fundamental goal of statistical modeling is to use the simplest model possible that still explains the data. This is known as principle of parsimony [18].

A model for a Stochastic time series is always called a Stochastic process and can be said to be a random variables family indexed by time (i.e., X_1, X_2, …) or generally (X_t) in discrete time space.

More precisely, $\{X_t, t \in T\}$ where T is the index of times on which the process is defined. The notation is necessary when observations are not equally spaced through time, but we restrict attention to the equally spaced case when the index set consisting of positive integers is commonly used [20].

2.1 Time series plot: - The time plot is the graphical representation of data. The first step in any time series analysis process is to plot the observed variables against time. The time plot reveals the presence of the likely component in the data.

2.2 Box-Jenkins Methodology

The general model introduced by [13] includes autoregressive as well as moving average parameters, and explicitly includes differencing in the formulation of the model. Specifically, the 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*); so, for example, a model described as (0, 1, 2) means that it contains 0 (zero) autoregressive (*p*) parameters and 2 moving average (*q*) parameters which were computed for the series after it was differenced once. The steps in Box-Jenkins methodology is as follows-

-Model identification

-Model estimation

-Model checking

2.2.1 Model Identification Phase

Before the estimation can begin, we need to identify the specific number and type of ARIMA parameters to be estimated. The major tools used in the identification phase are plots of the series, correlograms of auto correlation (ACF). The decision is not straightforward and in less typical cases requires not only experience but also a good deal of experimentation with alternative models (as well as the technical parameters of ARIMA).

However, a majority of empirical time series patterns can be sufficiently approximated using one of the 5 basic models that can be identified based on the shape of the autocorrelogram (ACF).

2.2.2 Model Checking

This phase involves checking for adequacy of the model, considering the properties of the residual. An overall check of model adequacy is provided by the Ljung-Box statistic (Q^*)

$$Q^* = T(T+2)\sum_{j=1}^{p} \frac{r_j^2}{T-j} \quad \text{Distributed } \chi^2_{\text{p-r}}$$

Where r_j^2 = residual autocorrelation at lag j

T = number of residual

P = number of time lags in the test

If the p-value associated with Q^* statistic is small (i.e. p < α), the model is inadequate.

We can consider modify or consider a new model until a satisfactory model is determined. The properties of the residual can be check using the following:

I. One-sample Kolmogorov-Smirnov test (check for normality considering normal probability plot or p = value)
II. Considering the graph of ACF and PACF of the residual. The individual residual autocorrelation should be small.

2.3 Autocovariance and Autocorrelation Function

The autocovariance of lag k is denoted as

$Y_k = e_k = \frac{1}{N-K}\sum(X_K - \bar{X}_K)(X_{K+1} - \bar{X}_K)$

Where k = 0, 1, 2,……

The autocorrelation lag k is obtain by dividing the autocovariance function of lag k by that of lag 0

ie, $e_k = \frac{ek}{eo}$

2.3.1 Autoregressive Model (AR)

The notation AR (*p*) refers to the autoregressive model of order *p*.

The AR (*p*) model is written as

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + £_t$$

Where $\varphi_1 \ldots \varphi_p$ are the parameters of the model, *c* is a constant and ε_t is white noise [10].

Likewise, it can be written as-

$Y_t = \theta_0 + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \ldots\ldots + \theta_p Y_{t-p} + £_t$

Where-

Y_t = response variable at (t)

Y_{t-1}, Y_{t-2} , ……. Y_{t-p} = response at t-1, t-2, … t-p respectively.

θ_0, $\theta_1 \ldots\ldots \theta_p$ = coefficient to be estimated

$£_t$ = Error term

The parameters of an autoregressive model can be estimated by minimizing the sum of squares residual with respect to each parameter [21].

2.3.2 Moving Average Model

In time series analysis the moving average (MA) model is a common approach for modeling univariate time series models. The notation MA (q) refers to the moving average model of order q:

$$X_t = \mu + £_t + \theta_1 £_{t-1} + \ldots + \theta_q £_{t-q} \quad .$$

In the compact form it can be re-written as:

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}$$

Where μ is the mean of the series, the $\theta_1 \ldots \theta_q$ are the parameters of the model and the $\varepsilon_t, \varepsilon_{t-1}, \cdots$ are error terms. The value of q is called the order of the MA model [22].

That is, a moving average model is conceptually a linear regression of the current value of the series against previous (unobserved) white noise error terms or random shocks. The random shocks at each point are assumed to come from the same distribution, typically a normal distribution, with location at zero and constant scale. The distinction in this model is that these random shocks are propagated to future values of the time series. Fitting the MA estimates is more complicated than with autoregressive models (AR models) because the error terms are not observable. This means that iterative non-linear fitting procedures need to be used in place of linear least squares. MA models also have a less obvious interpretation than AR models.

Sometimes the autocorrelation function (ACF) and partial autocorrelation function (PACF) will suggest that a MA model would be a better model choice and sometimes both AR and MA terms should be used in the same model [23].

2.3.3 Autoregressive Moving Average Model

The notation ARMA (p, q) refers to the model with p autoregressive terms and q moving average terms. This model contains the AR (p) and MA (q) models expressed as:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-q} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-q}$$

The error terms $£_t$ are generally assumed to be independent identically-distributed random variables (i.i.d.) sampled from a normal distribution with zero mean: $£_t \sim N(0, \sigma^2)$ where σ^2 is the variance [22].

An advantage of using an ARMA process to model a time series data is that an ARMA may adequately model a time series with fewer parameters, than using only an MA process or an AR process.

One fundamental goal of statistical modeling is to use the simplest model possible that still explains the data. This is known as principle of parsimony [5].

Likewise, it can be written as-

$Y_t = \theta_0 + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \ldots + \mathbf{Є_t} - \theta_1 Є_t - \theta_2 Є_{t-1} - \theta_2 Є_{t-2} - \ldots . - \theta_q Є_{t-q}$

Y_t = Response (dependent) variable at time t

θ_0 = constant mean

$\theta_1, \theta_2 \ldots . \theta_q$ = coefficients to be estimated

$Є_t$ = error terms at time t

$Є_{t-1}, Є_{t-2} \ldots . . Є_{t-q}$ = errors in the previous time periods that are incorporated in Y_t

2.3.4 Autoregressive Integrated Moving Average (ARIMA)

The general model introduced by [3] includes autoregressive as well as moving average parameters, and explicitly includes differencing in the formulation of the model. Specifically, the 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); so, for example, a model described as (1, 1, 2) means that it contains 1 (one) autoregressive (p) parameters and 2 moving average (q) parameters which were computed for the series after it was differenced once.

The input series for ARIMA needs to be stationary, that is, it should have a constant mean, variance, and autocorrelation through time. Therefore, usually the series first needs to be differenced until it is stationary (this also often requires log transforming the data to stabilize the variance). The number of times the series needs to be differenced to achieve stationarity is reflected in the d parameter.

In order to determine the necessary level of differencing, one should examine the plot of the data and autocorrelogram. Significant changes in level (strong upward or downward changes) usually require first order non seasonal (lag=1) differencing; strong changes of slope usually require second order non seasonal differencing. Seasonal patterns require respective seasonal differencing. If the estimated autocorrelation coefficients decline slowly at longer lags, first order differencing is usually needed. However, one should keep in mind that some time series may require little or no differencing, and that over differenced series produce less stable coefficient estimates.

At this stage which is usually called Identification phase, we also need to decide how many autoregressive (p) and moving average (q) parameters are necessary to yield an effective but still parsimonious model of the process. Parsimonious means that it has the fewest parameters and greatest number of degrees of freedom among all models that fit the data. In practice, the numbers of

the *p* or *q* parameters very rarely need to be greater than 2 [13].

3. Data Analysis

The data used in this study are the monthly numbers of road accident from January 2004 through December 2011. There are 96 data points employed. The data are collected from the National Headquarter of the Federal Road Safety Corps (FRSC),
Abuja.

3.1 Time Plot

Time plot which is the first step in data analysis is plotted. i.e. the graph of the original data versus time. The plot is given below:

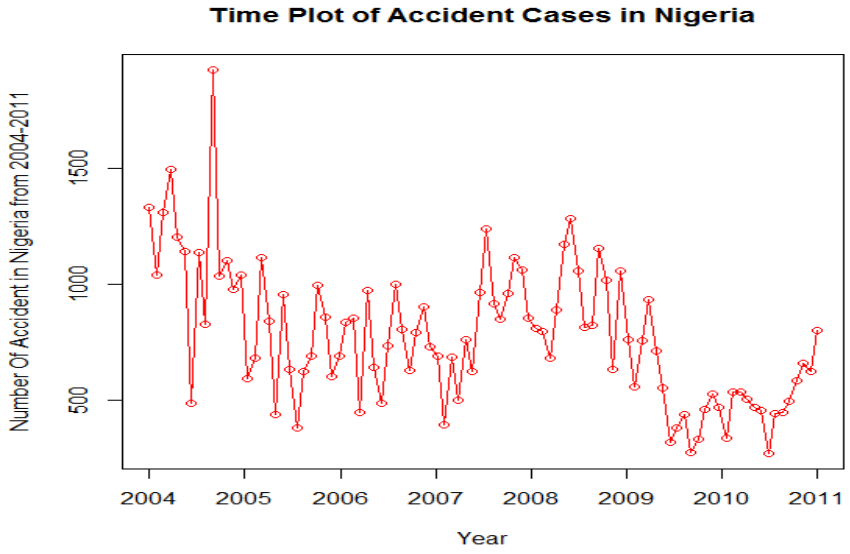

Fig 1: Time Plot

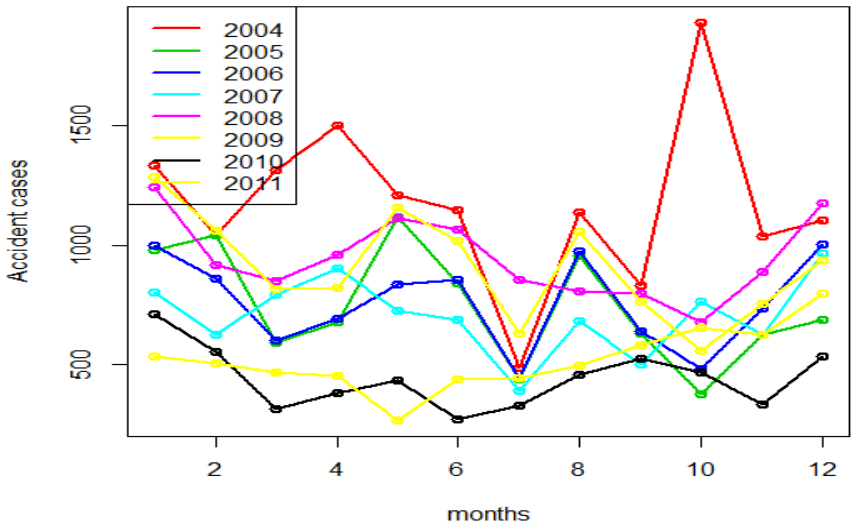

Fig 2: Time plot of each Year

3.2 Fitting of ARIMA model on Accident Data

3.2.1 First step: Model identification (to check for stationarity)

The Graph ACF is used to know whether a series is stationary or not. If the ACF graph of a time series values either cuts off fairly quickly or dies down fairly quickly, then the time series values should be considered Stationary. If the ACF graph dies down extremely slowly, then the time series values should be considered non-stationary.

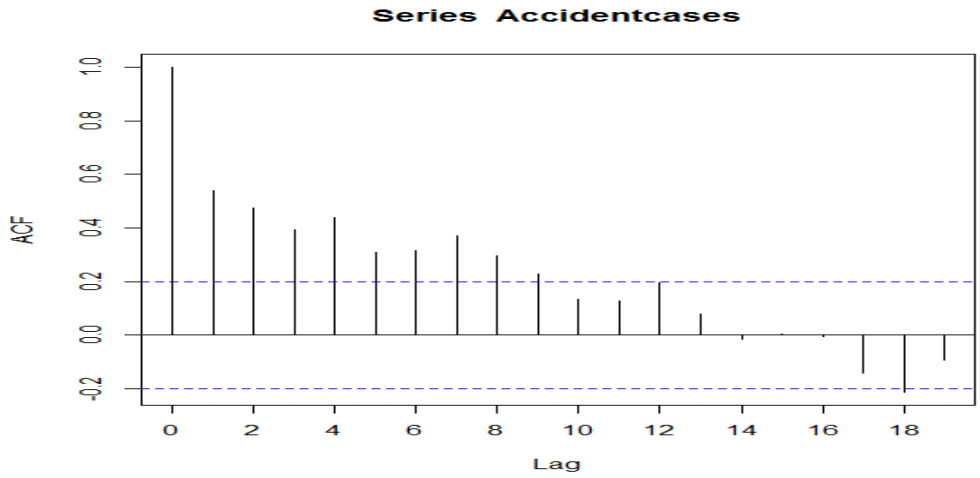

Fig 3: Graph of Autocorrelation Function of Accident data.

Observation: from Fig 4, it is obvious that the graph of ACF dies down extremely slowly, and the time series values should be considered non-stationary.

3.2.2 Test for stationarity of the data using Dickey-Fuller T statistic

Hypothesis

H_0: the data is not stationary (i.e. the data need not to be differenced to make it stationary)
Vs
H_1: the data is stationary (i.e. the data need to be differenced at least once to make it stationary)

Test statistic

Dickey -Fuller t statistic

DICKEY FULLER TEST
adf.test(Accidentcases,k=1)
Augmented Dickey-Fuller Test
data: Accident cases
Dickey-Fuller = -6.2842, Lag order = 1, p-value = 0.01

Alternative hypothesis: stationary
(Warning message:In adf.test(Accidentcases, k = 1) : p-value smaller than printed p-value)

Table 1: Dickey-Fuller t statistic result table

Decision rule: reject H_0 if the P-value < α=0.05, otherwise do not reject

Decision: since P-value(0.01) < α=0.05, we therefore reject H_0

Conclusion: we conclude therefore that the data is stationary after the 1st differenced.

Time Series Plot of the Difference 1

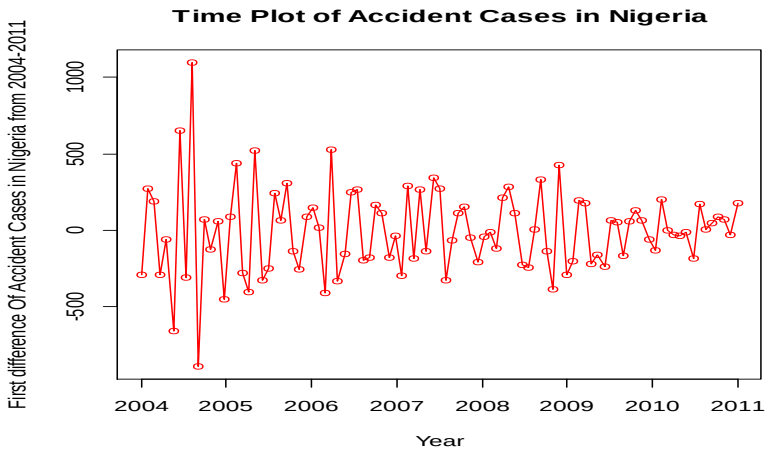

Fig 4: Time series plot of the first difference

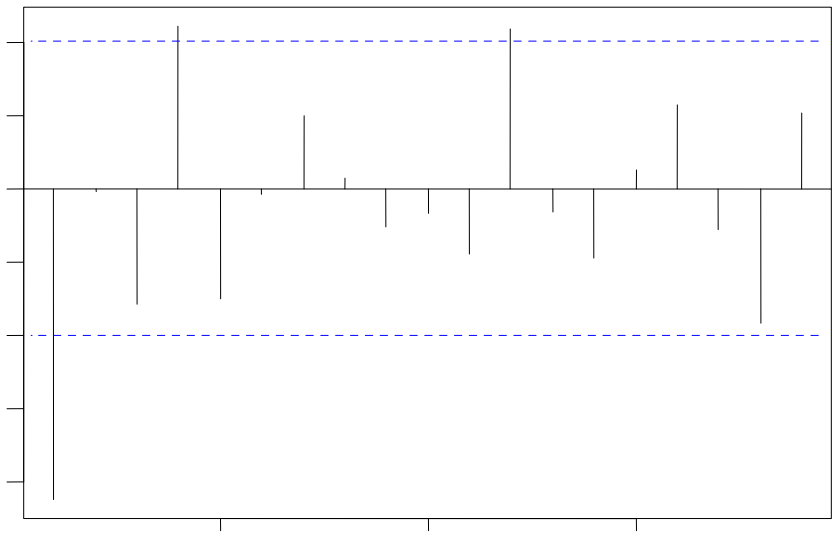

Fig 5: ACF plot of the first differenced data

Observation: from Fig 5, it is noticeable that the graph of ACF of the time series values cuts off quickly, then the times series is considered stationary at difference 1.

3.3 Second Step: Model Parameter Estimation

The parameters to be used are:-

- **ARIMA (AUTO-REGRESSIVE INTERGARTED MOVING AVARAGE)**
- **M.A (MOVING AVERAGE)**

Hypothesis testing:

H_0: the model is not significant

H_1: the model is significant

Decision rule: reject H_0 if P-value < 0.05

ARIMA (Auto-Regressive Integrated Moving Average Model)

Estimates of Parameters of ARIMA (1, 1, 1)

Type	Coef	SE Coef	T	P
AR1	-0.4199	0.0935	-4.4893	2.067e-05
MA1	-1.000	0.029	-34.4925	<2.2e-16
Differencing:1	Number of observations: Original series 96		after differencing 95	
Residuals:	MS= 64588	DF = 93		
AIC	1317.33			

Table 2: table of parameter estimate for order 1 (ARIMA)

Model for ARIMA (1, 1, 1) is giving by: $\hat{Y}_t = -0.4199Y_{t-1} - 1.000e_{t-1}$

From Table 2, comparing the P-value estimated with the α-value; the ARIMA(1,1,1) parameters estimated are significant. The MSE & AIC estimated values are 64588 & 1317.33 respectively.

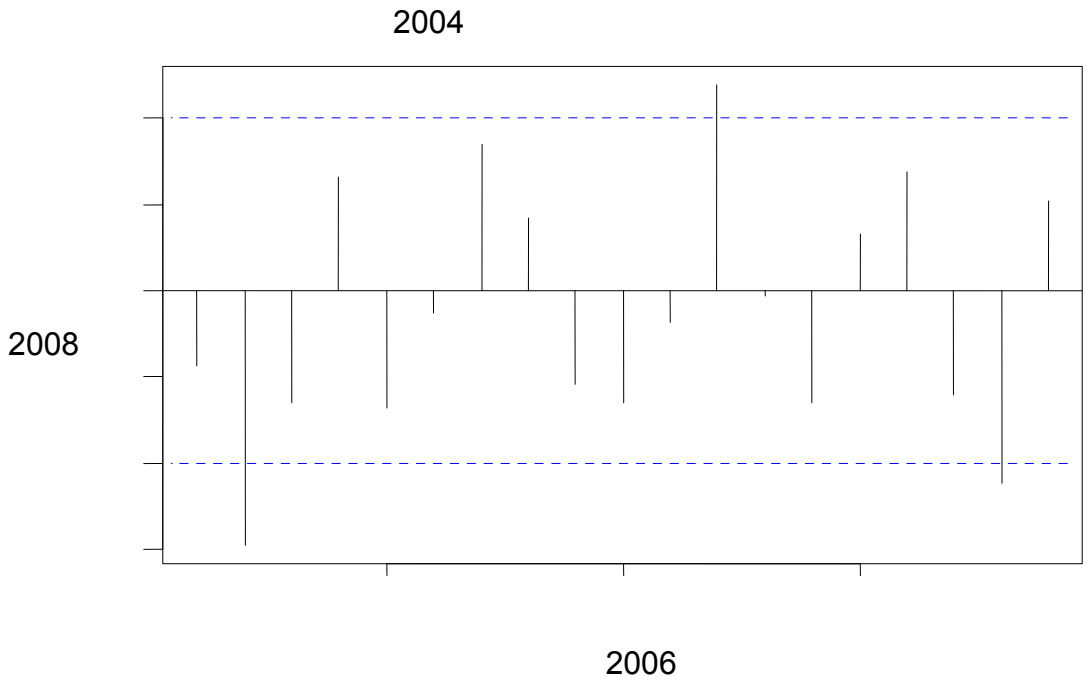

Fig 6: Plot for ARIMA (1, 1, 1)

Observations: from fig 6, ACF for residuals are significant at some lag (2,12, 18), meaning that serial correlation is significant between the error terms i.e. the model is not adequate.

Estimates of Parameters of ARIMA (2, 1, 1)

Type	Coef	SE Coef	T	P
AR1	-0.5118	0.1009	-4.4893	2.063e-06
AR2	-0.2139	0.1007	-2.1233	0.03645
MA1	-1.000	0.0307	-32.5832	<2.2e-16
Differencing:1	Number of observations: Original series 96		after differencing 95	
Residuals:	MS= 61338	DF = 92		
AIC	1314.95			

Table 3: table of parameter estimate for order 2 (ARIMA)

Model for ARIMA (2, 1, 1) is giving by: $\hat{Y}_t = -0.5118Y_{t-1} - 0.2139Y_{t-2} - 1.000e_{t-1}$

From table 3, comparing the P-value estimated with the α-value; the ARIMA(2,1,1) parameters estimated are significant. The MSE & AIC estimated values are 61338 & 1314.95 respectively.

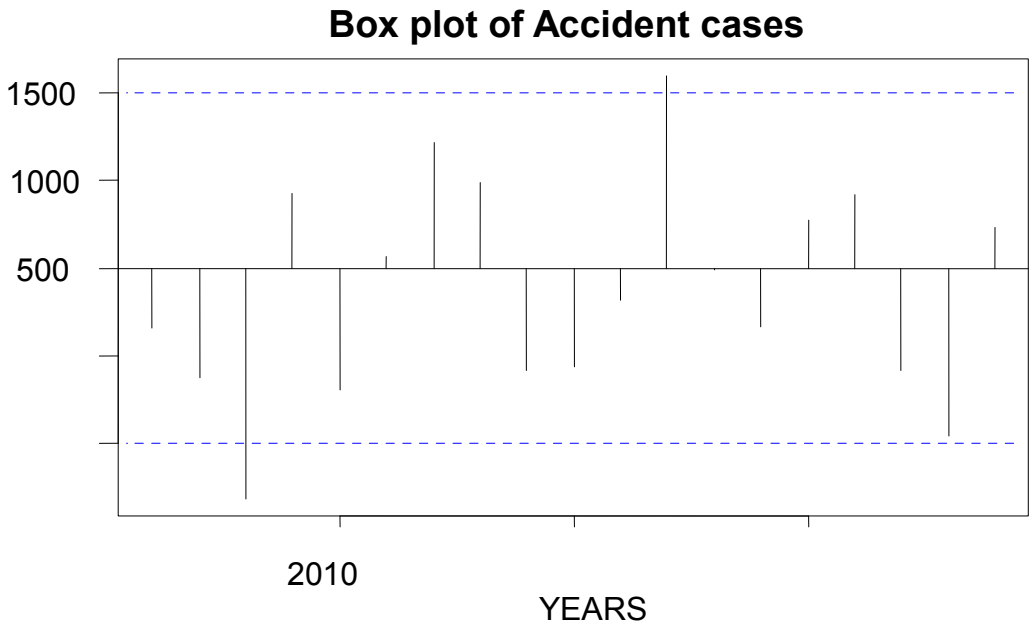

Fig 7: Plot for ARIMA (2, 1, 1)

Observations: from fig 7, ACF for residuals are significant at some lag (3,12), meaning that serial correlation is significant between the error terms i.e the model is not adequate

Estimates of Parameters of ARIMA (3, 1, 1)

Type	Coef	SE Coef	T	P
AR1	-0.5855	0.0976	-5.995	4.072e-08
AR2	-0.3853	0.1087	-3.5635	0.0005879
AR3	-0.3195	0.0975	-3.2771	0.0014910
MA1	-1.000	0.0382	-26.1876	<2.2e-16
Differencing:1	Number of observations: Original series 96		after differencing 95	
Residuals:	MS= 54617	DF = 91		
AIC	1306.93			

Table 4: table of parameter estimate for order 3 (ARIMA)

Model for ARIMA (3, 1, 1) is giving by: $\hat{Y}_t = -0.5855Y_{t-1} - -0.3853Y_{t-2} - -0.3195Y_{t-3} - 1.000e_{t-1}$

From table 4, comparing the P-value estimated with the α-value; the ARIMA (3,1,1) parameters estimated are significant. The MSE & AIC estimated values are 54617 & 1306.93 respectively

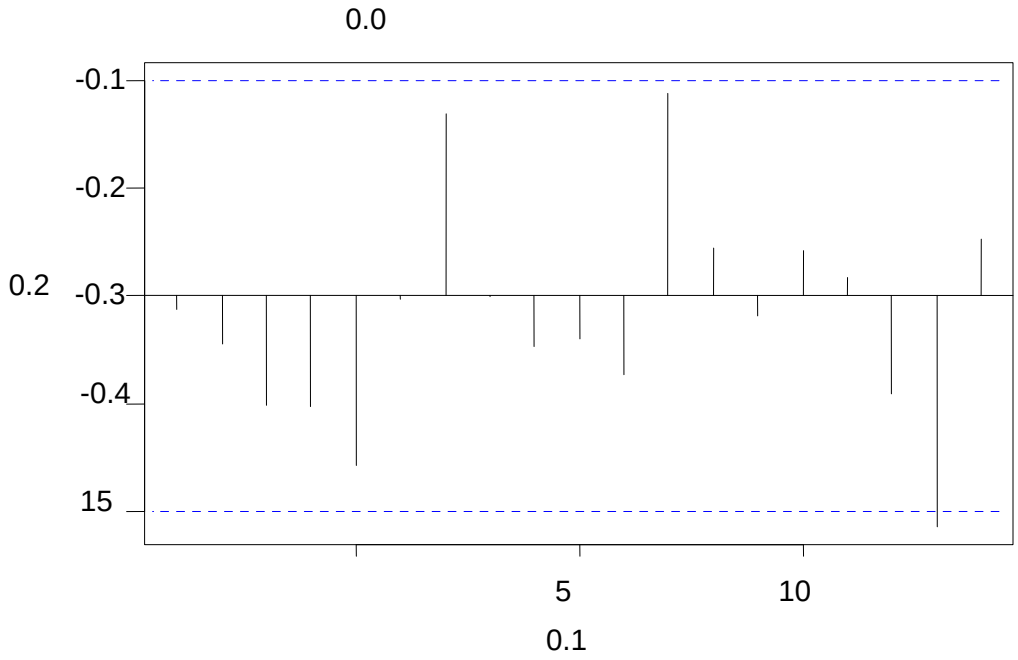

Fig 8: Plot for ARIMA (3, 1, 1)

Observations: from fig 8, ACF for residuals is significant at lag 18; meaning that serial correlation is significant between the error terms. Considering lag 1-5, the model is adequate

Estimates of Parameters of ARIMA (4, 1, 1)

Type	Coef	SE Coef	T	P
AR1	-0.5896	0.1042	-5.6588	1.832e-07
AR2	-0.3900	0.1159	-3.3647	0.001132
AR3	-0.3266	0.1160	-2.8142	0.006018
AR4	-0.0118	0.1040	-0.1132	0.910105
MA1	-1.0000	0.038		
Differencing:1	Number of observations: Original series 96		after differencing 95	
Residuals:	MS= 54595	DF = 90		
AIC	1308.92			

Table 5: table of parameter estimate for order 4 (ARIMA)

Model for ARIMA (4, 1, 1) is giving by:

$\hat{Y}_t = -0.5896Y_{t-1} - 0.3900Y_{t-2} - 0.3266Y_{t-3} - 0.0118Y_{t-4} - 1.0000e_{t-1}$

From table 5, comparing the P-value estimated with the α-value; the AR (4) parameters estimated is not significant. The MSE & AIC estimated values are 54595 & 1308.92 respectively

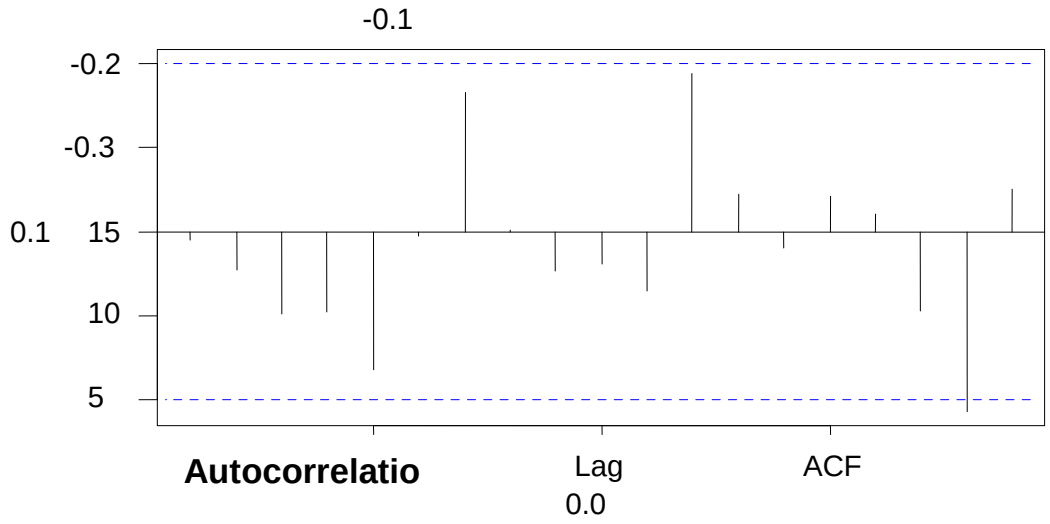

Fig 9: Plot for ARIMA (4, 1, 1)

Observations: from fig 9, ACF for residuals are significant at some lag 18, meaning that serial correlation is significant between the error terms. Considering lag 1-5, the model is adequate

Estimates of Parameters of ARIMA (5, 1, 1)

Type	Coef	SE Coef	T	P
AR1	-0.5950	0.1032	-5.7676	1.17e-07
AR2	-0.4372	0.1209	-3.6164	0.0004975
AR3	-0.3795	0.1225	-3.0972	0.0026212
AR4	-0.0904	0.1210	-0.7467	0.4572094
AR5	-0.1259	0.1025	-1.2279	0.2227666
MA1	-1.0000	0.0443	-2.5554	<2.2e-16
Differencing:1	Number of observations: Original series 96		after differencing 95	
Residuals:	MS= 53550	DF = 89		
AIC	1309.43			

Table 6: table of parameter estimate for order 5 (ARIMA)

M.A (MOVING AVERAGE MODEL)

Estimates of Parameters of MA (1)

Type	Coef	SE Coef	T	P
MA1	-1.000	0.0275	-36.336	<2.2e-16
Differencing:1	Number of observations: Original series 96		after differencing 95	
Residuals:	MS= 78997	DF = 94		
AIC	1333.37			

Table 7: table of parameter estimate for order 1 (MA)

Model for MA (1) is giving by: $\hat{Y}_t = -1.0000e_{t-1}$

From table 7, comparing the P-value estimated with the α-value; the MA(1) parameter estimated is significant. The MSE & AIC estimated values are 78997 & 1333.37 respectively.

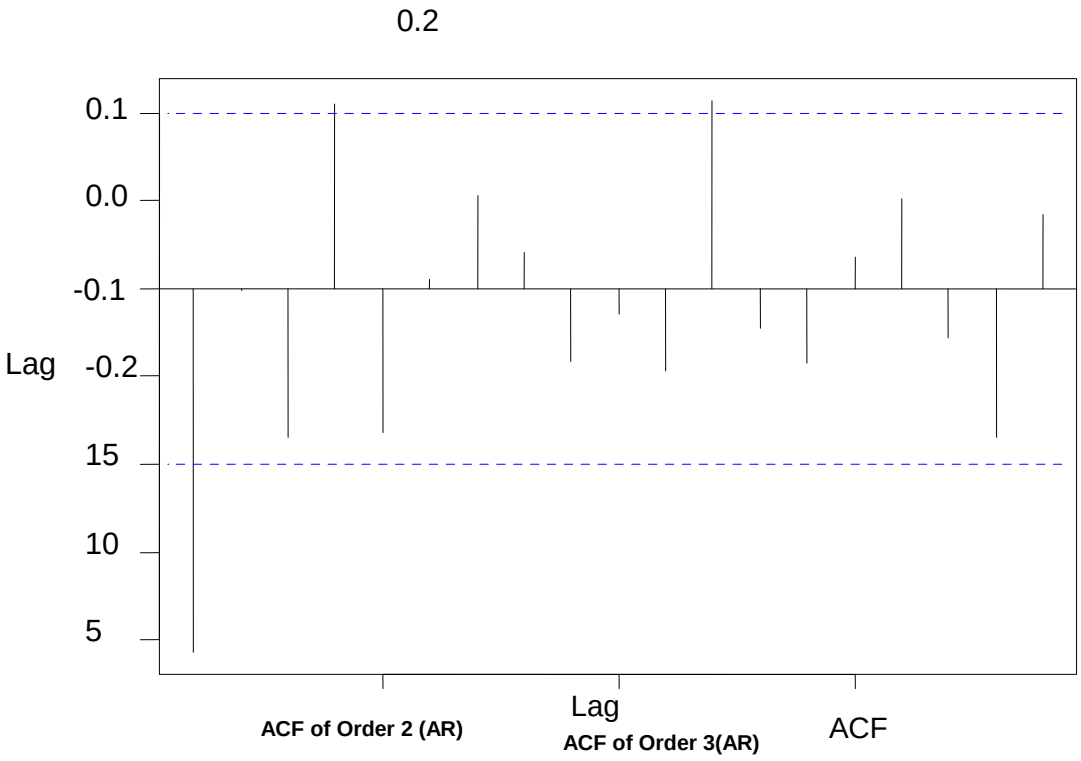

Fig 10: Plot for MA (0, 1, 1)

Observations: from fig 10, ACF for residuals are significant at some lag (1,12), meaning that serial correlation is significant between the error terms i.e the model is not adequate.

Estimates of Parameters of MA (2)

Type	Coef	SE Coef	T	P
MA1	-1.6944	0.0963	-17.5951	<2.2e-16
MA2	0.6944	0.0891	7.7916	9.746e-12
Differencing:1	Number of observations: Original series 96		after differencing 95	
Residuals:	MS= 55050	DF = 93		
AIC	1306.4			

Table 8: table of parameter estimate for order 2 (MA)

Model for MA (2) is giving by: $\hat{Y}_t = -1.6944e_{t-1} + 0.6944e_{t-2}$

From table 8, comparing the P-value estimated with the α-value; the MA(2) parameter estimated is significant. The MSE & AIC estimated values are 55050 & 1306.4 respectively.

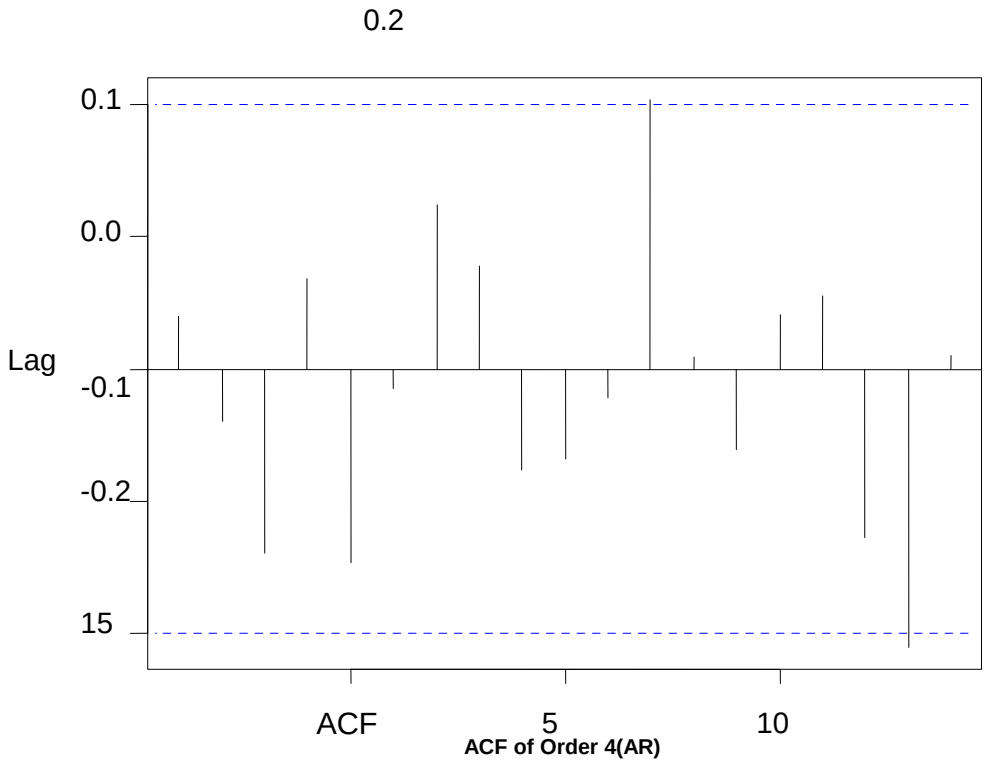

Fig 11: Plot for MA (0, 1, 2)

Observations: from fig 11, ACF for residuals are significant at lag (18), meaning that serial correlation is significant between the error terms. But considering lag 1-17, the model is adequate.

Estimates of Parameters of MA (3)

Type	Coef	SE Coef	T	P
MA1	-1.6520	0.1163	-14.2023	<2.2e-16
MA2	0.5933	0.1845	3.2157	0.001802
MA3	0.0587	0.0959	0.6121	0.542016
Differencing:1	Number of observations: Original series 96		after differencing 95	
Residuals:	MS= 54774	DF = 92		
AIC	1306.03			

Table 9: table of parameter estimate for order 3 (MA)

Model for MA (3) is giving by: $\hat{Y}_t = -1.6520e_{t-1} + 0.5933e_{t-2} + 0.0587e_{t-3}$

From table 9, comparing the P-value estimated with the α-value; the MA(3) parameter estimated is not significant. The MSE & AIC estimated values are 52774 & 1306.03 respectively.

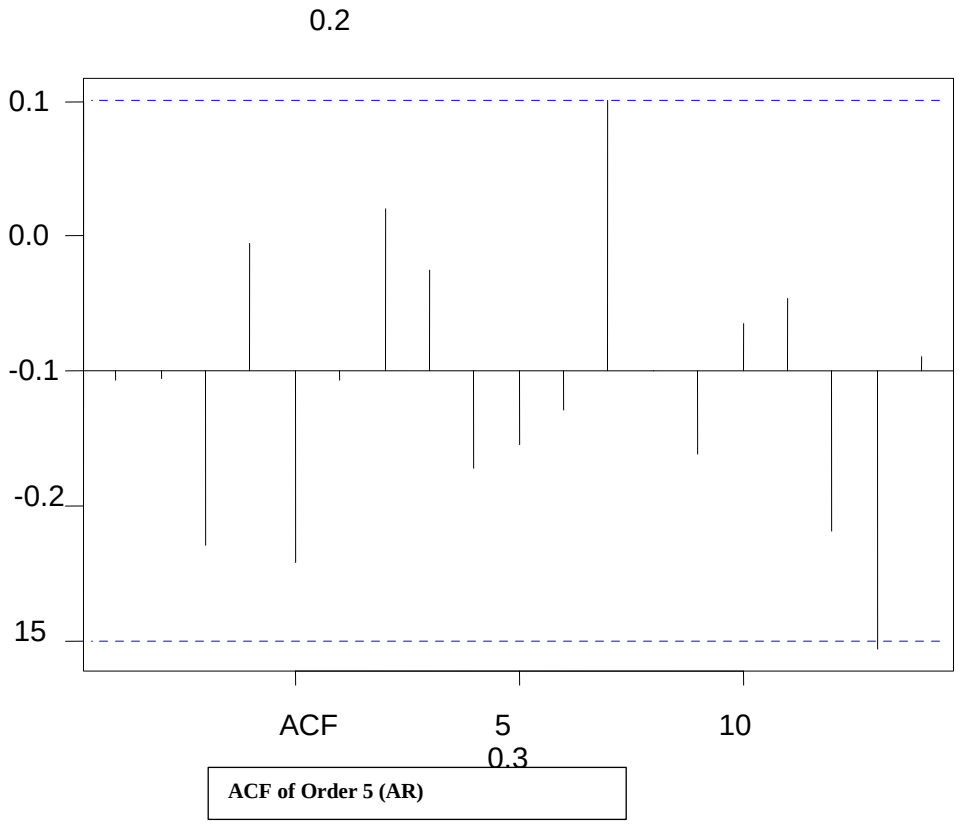

Fig 12: Plot for MA (0, 1, 3)

Observations: from fig 12, ACF for residuals are significant at lag (18), meaning that serial correlation is significant between the error terms. But considering lag 1-17, the model is adequate.

Estimates of Parameters of MA (4)

Type	Coef	SE Coef	T	P
MA1	-1.6471	0.1214	-135682	<2.2e-16
MA2	0.595	0.188	3.1654	0.002114
MA3	0.0361	0.1666	-5.6431	0.828782
MA4	0.0160	0.0964	-2.5712	0.868600
Differencing:1	Number of observations: Original series 96		after differencing 95	
Residuals:	MS= 54742	DF = 91		
AIC	1308			

Table 10: Table of parameter estimate for order 4 (MA)

Model for MA (4) is giving by: $\hat{Y}_t = -1.6471e_{t-1} + 0.595e_{t-2} + 0.0361e_{t-3} + 0.0160e_{t-4}$

From table 10, comparing the P-value estimated with the α-value; the MA(4) parameter estimated is not significant. The MSE & AIC estimated values are 54742 & 1308 respectively.

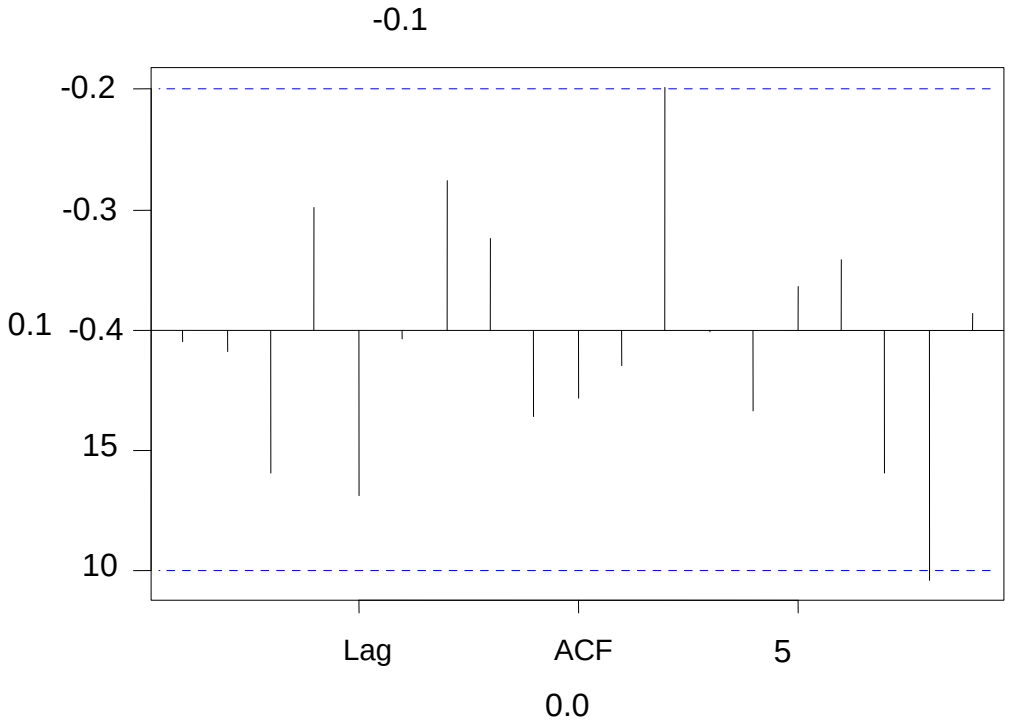

Fig 13: Plot for MA (0, 1, 4)

Observations: from fig 13, ACF for residuals are significant at lag (18), meaning that serial correlation is significant between the error terms. But considering lag 1-17, the model is adequate.

Estimates of Parameters of MA (5)

Type	Coef	SE Coef	T	P
MA1	-1.5986	0.1226	-13.0383	<2.2e-16
MA2	0.4753	0.2079	2.2858	0.02464
MA3	-0.0176	0.1662	-0.1056	0.91615
MA4	0.3427	0.2133	1.6068	0.11164
MA5	-0.2018	0.1203	-1.6768	0.09710
Differencing:1	Number of observations: Original series 96		after differencing 95	
Residuals:	MS= 53287	DF = 90		
AIC	1307.28			

Table 11: table of parameter estimate for order 5 (MA)

Model for MA (5) is giving by: $\hat{Y}_t = -1.5986e_{t-1} + 0.4753e_{t-2} - 0.0176e_{t-3} + 0.3427_{t-4} - 0.2018_{t-5}$

From table 11, comparing the P-value estimated with the α-value; the MA(5) parameter estimated not significant. The MSE & AIC estimated values are 53287 & 1307 respectively.

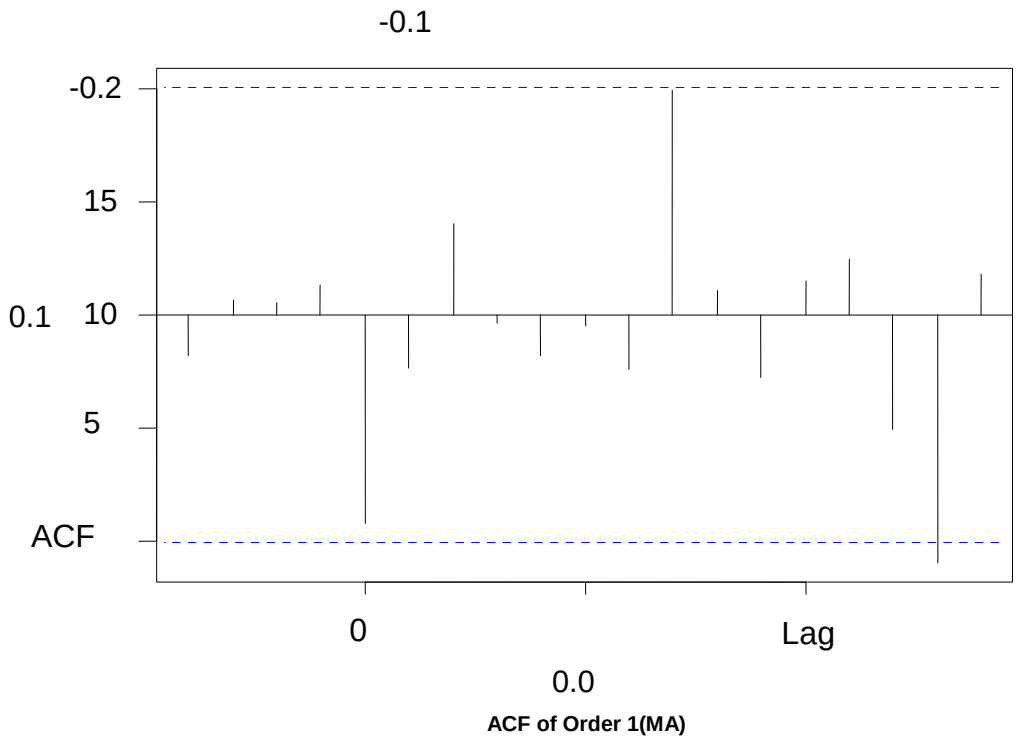

Fig 14: Plot for MA (0, 1, 5)

Observations: from fig 19, ACF for residuals are significant at lag (18), meaning that serial correlation is significant between the error terms. But considering lag 1-17, the model is adequate.

4. Summary, Discussion and Interpretation

4.1 Summary of Result Table

The summary of result table for the values at different orders is given below

Model	Significant status	ACF status	MSE	AIC
ARIMA(1,1,1)	Significant	Not adequate	64588	1317.33
ARIMA(2,1,1)	Significant	Not adequate	61338	1314.95
ARIMA(3,1,1)**	Significant	Adequate	54617	1306.93
ARIMA(4,1,1)	Not significant	Adequate	54595	1308.92
ARIMA(5,1,1)	Not significant	Adequate	53550	1309.43
MA(0,1,1)	Significant	Not adequate	78997	1333.37
MA(0,1,2)**	Significant	Adequate	55050	1304.4
MA(0,1,3)	Not significant	Adequate	54774	1306.03
MA(0,1,4)	Not significant	Adequate	54742	1308
MA(0,1,5)	Not significant	Adequate	53287	1307.28

Table 12: Summary of results
****** is selected as the best models for the data

4.2 Discussion of Results

From the time plot of the raw data in Fig 1, it could be seen that the highest number of accident occurred in October 2004 and the lowest total number of accident occurred in the year May 2011. The time plot indicated a non stationary series and the stationary series were obtained by taking the first difference of the original accident data. In fig 2, it shows that number of accidents is regularly high from October to January each year and usually low in July every year. Considering

summary statistics in Fig 3, shows that dispersion is low in 2007.

From table 11, models of ARIMA and MA were estimated at order 1-5 each. Test was conducted to know the significance level of each model, (i.e, to know which parameter contributes significantly to the model). From the analysis, we discovered that ARIMA(1,1,1), ARIMA(2,1,1), ARIMA(3,1,1), MA(0,1,1) and MA(0,1,2) are significant, which means that they contribute significantly to the model (are the good models for the data).

Furthermore, study was done on the Autocorrelation function graph for residuals (ACF), to know if the serial correlation between the error terms is significant or not. It was observed that the ACF for residuals are not significant on some models which make them adequate (i.e. the serial correlation between the error terms is not significantly different from zero (0)). The models that are adequate are ARIMA(3,1,1), ARIMA(4,1,1), ARIMA (5,1,1), MA (0,1,2), MA (0,1,3), MA(0,1,4), MA(0,1,5).

Mean Square Error (MSE) and Akaike Information Criteria was also employed in selection of the best model. Considering the significant models ARIMA (3,1,1) gave the lowest MSE value and MA (0,1,2) gave the lowest AIC value.

Therefore, from table 11, the models; ARIMA (3,1,1) and MA (0,1,2) were discovered to be Significant and their ACF for residual are not significant with low MSE and AIC values respectively. Hence, ARIMA (3,1,1) and MA (0,1,2) were the best model for fitting and forecasting road accident data in Nigeria.

4.3 Interpretation of results

It was discovered that ARIMA (3,1,1) and MA (0,1,2) were the best model for road accident data in Nigeria-

- **ARIMA (3,1,1) model is express as:-**

$$\hat{Y}_t = -0.5855Y_{t-1} - 0.3853Y_{t-2} - 0.3195Y_{t-3} - 1.000e_{t-1}$$

Where-

Y_t= number of road accident in the projected month

Y_{t-1} = number of road accident of the immediate past month

Y_{t-2}= number of road accident before the immediate past month

Y_{t-3} = number of road accident of the month preceding the accident before the immediate past month (i.e. last 2 month away)

e_{t-1}= estimated error in the immediate past month.

- **MA (0,1,2) model is express as:-**

$$\hat{Y}_t = -1.6944e_{t-1} + 0.6944e_{t-2}$$

Y_t = number of road accident in the projected month

e_{t-1}= errors in the immediate past month

e_{t-2}= error in the month preceding the last month

5. Conclusion

From the time plot, it can be shown that the number of road accidents between the year 2004-2011 do not follow a particular trend (upward or downward trend) but in the recent years, downward trend was being experienced, this can be as a result of intensified efforts of the Road Safety Corps whose vision is to eradicate road accident and create safe motoring environment in Nigeria.

It was also discovered that road accidents is always low in July and always high between October to January each year. Considering dispersion of the accident data, it was discovered to be low in 2007.

Also from the discussion above, ARIMA (3,1,1) and MA (0,1,2) models best fit these data collected for forecasting purposes and policy formation.

6. Recommendation

From the analysis conducted in this research work and outcome of our findings, we decide to offer these responsive recommendations for the stakeholders in Nigeria:

(1) The Federal Road Safety Corp should upgrade their effort more in term of sensitization of the road users on the rules guiding driving and provide Severe punishment for road law offenders

(2) Due to apparent increasing trend in the outcome of road accident on our road, the government should look into the poor state of the country's road being a major cause of road accident.

(3) More efforts should be concentrated on the maintenance of our road as is being championed by FERMA.

References

[1] Spiegel M.R, Stephens L.J (1999), Theory and Problems of Statistics (Third edition)

[2] Frankowne and Ronjones (1982), "Statistics" 2nd Edition, Pitman, Chapter 14 "Sample Design" and Chapter 15, "Planning a Sample Survey"

[3] Chatfield (1987), The Analysis of Time Series: An Introduction. London Chapman and Hall (Third Edition)

[4] Harvey A.C.and R.D. Synder (1990), Structural Time Series Models in Inventory Control, International Journal of Forecasting, 6, 187-198.

[5] McCleary, R. and Hay, R.A., Jr (1980), Applied Time Series Analysis for the Social Sciences, Beverly Hills, CA: Sage.

[6] Yule, G.U. (1927), On a Method of Investigating Periodicities in Distributed series with Special Reference to Wolfer's Sun spot Numbers. Philosophical Transition Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character, 226: 267-298.

[7] Kolmogorov, A.N. (1941), Stationary Sequences in Hilbert space (in Russian). Bull. Math. Univ. Moscow, 2(6), 1-40.

[8] Newbold, P. (1983), ARIMA Model building and Time Series Analysis approach to Forecasting, Journal of Forecasting, 2, 23-35.

[9] Box George and Jenkins, Gwilyn (1970), Times Series Analysis, Forecasting and Control, Holden Day, San Francisco.

[10] Yule, G.U. (1926), "Why Do We Sometimes Get Nonsense-Correlation between Time Series? A study

In sampling and the Nature of Time Series, Journal of Royal Statistical Society, 89, 1-64.
[11] Slutsky, E., (1937), "The Summation of Random Causes as the Source of Cyclic Processes". Econometrica 5, 105-146.
[12] Wold, H. (1938), A study in the Analysis of Stationary Time Series. Almgrist and Wiksell, Stockhom.
[13] Box George & Jenkins, Gwilyn (1976), Time Series Analysis: Forecasting and Control. San Francisco:Holden Day.
[14] Cooper, R.L. (1972), "The Predictive Performance of Quarterly Econometric Models of the United States", In Econometric Models of Cyclical Behavior (B.G. Hickman, Ed) New York: National Bureau of Economic Research.
[15] Nelson, C.R. (1972), "The Prediction Performance of FRB-MIT-PENN Model of the US Economy",American Economic Review, 5, 902-917.
[16] Elliot, J.W. (1973), "A Direct Comparison of Short-Run GNP Forecasting Models", Journal of Business, 46, 33-60.
[17] Narasimham et al (1974), "On the Predictive Performance of the BEA Quarterly Econometric Model andBox-Jenkins Type ARIMA Model". Proceedings of the American Statistical Association: Business and Economics section, pp. 501-504.
[18] McWhhorter, A. Jr. (1975), "Time Series Forecasting using the Kalman Filter: An Empirical Study", Proceeding of American Statistical Association. Business and Economic Section, pp. 436-446.
[19] Armstrong, J.S. (1978), "Forecasting with Econometric method: Folklore versus Fact with Discussion", Journal of Business, 51, 549-600.
[20] Chatfield (1978), The Analysis of Time Series: An Introduction. London Chapman and Hall (Sixth Edition)
[21] Brockwell P. J and Lindner A. (1987), Existence and Uniqueness of Stationary Levy-driven CARMA Processes Stoch. Proc. Appl. 119, 2260-2681.
[22] Harris, Richard and Robert, Sollis (2003), Applied Time Series Modeling and Forecasting, John Wiley and Sons, Chichester.
[23] Mills T. C. (1990), Time Series Techniques for Economists. Cambridge University Press.

Author Profile

Author Name: Balogun Oluwafemi Samson received the B.Sc. and M.Sc. degrees in Statistics from University of Ilorin in 2007 and 2010, respectively. He currently running his Ph.D. programme in Statistics in the same institution and he is now with Department of Statistics and Operations Research, Modibbo Adama University of Technology, Yola as an Academic Staff.