# Time Series Analysis of Road Accidents in Ghana

**Salifu Nanga,** *(M.Phil. MSc. BSc.)*
*Department of Statistics, School of Graduate Studies, University of Ghana,*

***Abstract***
*The mortality and morbidity trends from road accidents are rising annually in almost all the developed countries. The total number of deaths and injuries continues to rise despite the decrease in road casualties in relation to the total number of vehicles. Road traffic accident in Ghana is also increasing at an alarming rate and has raised major concerns. In this study, univariate time series was used to model road accidents. The Box – Jenkins method was applied for a 20 year period from 1991-2010.A model was subsequently developed to fit the time series data. SARIMA (1,1,0) × (0,1,1)$_{12}$ was found to be the best model for road accident cases with a maximum log likelihood value of 245.48, and least AIC value of 5.892, RMSE value of 17.930 and MAPE value value of 1.688. An ARCH-LM test and Ljung-Box test on the residuals of the models revealed that they are free from heteroscedasticity and serial correlation respectively.*

**Keywords:** *ARIMA model, SARIMA model, Bank of Ghana, Box Jenkins, Forecasting, Road Accidents*

## I. INTRODUCTION

Road accident is the eleventh most common cause of death in the world and accounts for 2.1% of all deaths globally (World Report on Road Traffic Injury Prevention, 2004).Ghana loses 1.6 % of Gross Domestic Product (GDP) through road accidents (NRSC, 2007). Between the short period from January to August 2011, 1,431 people have died as a result of road accidents (Ghana web, 2011). This is a very alarming phenomenon which has to be tackled tactfully. Some of the socio-economic effects include disability and therefore a high dependency burden, and for some victims the seriousness of their disability could result in being jobless. With men representing 70% of national casualties, it has serious implications.

Monthly data on road accidents were collected in Ogun State. The Box-Jenkins approach of model identification, parameter estimation and diagnostic check was adopted. The result revealed that the monthly road accidents is basically an ARIMA (2,1,1) model, ARIMA (2,1,0) model, ARIMA (2,1,1) model (Olobatuyi ,2012). A study carried out in Lagos state on road traffic accidents for the period 2005 – 2010 revealed that the ARIMA model of order (2, 1, 0) was appropriate (Fadugba ,2012). A study also carried out in Egypt revealed that the monthly series of the number of road traffic accidents from 1990 to 2008 was used, where results showed that the adequate model is SARIMA (1,1,1)*(0,0,1)$_{12}$ (Abass, 2004). Another study by CAPMAS for monthly data on road accidents Egypt revealed that SARIMA(1,1,1)*(0,1,1)$_{12}$ was appropriate for road accidents, SARIMA (0,1,1)*(0,0,2)$_{12}$ for fatalities and SARIMA (1,1,1)*(1,1,1)$_{12}$ (Ho et al , 2002).

To analyze annual road accident data between 1980 and 2010, a Box-Jenkins approach was used to determine the patterns of road traffic accident cases, injuries and deaths along the Accra-Tema motorway (Okutu, 2011). ARIMA (1, 1, 1) was used to model injury and death as a result of accidents while ARIMA (0, 1, 2) was used to model accident cases. Forecasts revealed that there will be a steady increase in road accidents in the next five years.

A statistical analysis of the systematic yearly increase in the number of accidents was carried out by Boakye et al in 2013. Data was collected on yearly road traffic accidents and population values of Ghana for the period 1990 to 2012.The study revealed a trend that was systematic in pattern of growth in both road traffic accidents and population; the study also revealed a strong a strong positive correlation existed between road traffic accidents and population.

A similar study was carried out in Ghana by Adu-Poku et al in 2014.ARIMA (0, 2, and 1) was identified as the best model. Road accidents normally occur during festive seasons (Mustafa, 2005). As a result there is the possibility of a seasonal pattern in the dataset. However they failed to test for seasonality in the data. In this study, the issue of seasonality will be addressed and a robust model to fit road accident data will be built.

## II. MATERIALS AND METHODS

This study was carried out in Ghana using monthly data which covered the period January, 1991 to December, 2010. Secondary data was collected from the Ghana Road Safety Commission. The data was modeled using Autoregressive Integrated Moving Average (ARIMA) stochastic model. An autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. These models are fitted to time series data either to better understand the data or to predict future points in the series.

### *A. Autoregressive Integrated Moving Average (ARIMA) Model*

A time series $Y_t$ is said to follow Autoregressive Integrated Moving Average (ARIMA) model if the *d*th differences $\nabla^d Y_t$ follow a stationary ARMA model. There are three important things that characterize an ARIMA building process (Tebbs, 2010):

- *p*, the order of the autoregressive component
- *d*, the number of differences needed to arrive at a stationary ARMA(*p*, *q*) process
- *q*, the order of the moving average component

The general form of the ARIMA (*p*,*d*,*q*) is represented by a backward shift operator as

$$\emptyset(B)(1-B)^d Y_t = \theta(B)e_t, \qquad (1)$$

where the AR and MA characteristic operators are

$$\emptyset(B) = (1 - \emptyset_1 B - \emptyset_1 B^2 - \text{---} - \emptyset_p B^d) \qquad (2)$$

$$\theta(B) = (1 - \theta_1 B - \theta_2 B^2 - \text{---} - \theta_q B^q) \qquad (3)$$

and

$$(1-B)^d Y_t = \nabla^d Y_t \qquad (4)$$

where

$\emptyset$ is the parameter estimate of the Autogressive component
$\theta$ is the parameter estimate of the Moving Average component
$\nabla$ is the difference
$B$ is the Backward shift operator
$e_t$ is a purely a random process with mean zero and var($e_t$) = $\sigma^2_e$

The estimation of an ARIMA model was first approached by Box and Jenkins (1976) and according to their methodology, as a result it involves three steps as Identification, Estimation, and Diagnostic Checking. The three steps can be summarized in the following below (Tebbs, 2010).

### *B. Seasonal ARIMA(p,d,q)(P,D,Q)s*

ARIMA models cannot really cope with seasonal behavior. It only models time series with trends. Seasonal ARIMA models are defined by 7 parameters ARIMA (p,d,q)(P,D,Q)s. A general definition of the Seasonal ARIMA models is

Where

- AR(p) Autoregressive part of order p
- MA(q)Moving average part of order q
- I (d) differencing of order d
- ARs (P) Seasonal Autoregressive part of order P
- MAs (Q) SeasonalMoving average part of order Q
- Is (D) seasonal differencing of order D
- s is the period of the seasonal pattern appearing i.e. s = 12 months data.

SARIMA (p, d, q) x (P, D, Q)s is given by $\phi(Bs)\phi(B)\ \nabla Ds\nabla dxt = \Theta(Bs)\theta(B)wt$

### *C. Model Identification*

Identification step involves the use of the techniques to determine the values of *p*, *q* and *d*. The values are determined by using Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). For any 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 the theoretical ACF has non zero autocorrelation at lags 1, 2, …, *q* and zero autocorrelations at all lags. The non-zero lags of the sample PACF and ACF are tentatively accepted as the p and q parameters. Bad choices of *p*, *d*, and *q* lead to bad models, which, in turn, lead to bad predictions (forecasts) of future values (Tebbs, 2010).

### *D. Unit Root Tests*

Determining whether the time series is stationary or not is a very important concept before making any inferences in time series analysis. Therefore Augmented Dickey Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin tests have been used to check the stationarity of the data series. The test is based on the assumption that a time series data $y_t$ follows a random walk (Mahadeva, 2004):

$$Y_t = \rho y_{t-1} + e_t \ldots\ldots\ldots\ldots\ldots (5)$$

where $\rho$ is the characteristic root of an AR polynomial and $e_t$ is purely a random process with mean zero and variance $\sigma^2$ (Tebbs,2010).

#### *a. Augmented Dickey Fuller (ADF) Test*

Dickey and Fuller (1979) ADF is a test to see if the test can reject non-stationarity. The ADF unit root test therefore tests [12],

$$H_0: \rho = 1 \text{ (non-stationary)}$$

versus
$H_1$: $\rho < 1$ (stationary)

#### *b. Kwiatkowski Phillips Schmidt Shin (KPSS) Test*

Kwiatkowski et al (1992) KPSS is a test where the null hypothesis is the other way around. It is tests to see if the test can reject stationarity. This is the reverse of ADF test (Mahadeva et al,2004).

### *E. Estimation of Model Parameters*

After identifying the possible ARIMA models, the maximum likelihood method is used to estimate the model parameters (Tebbs,2010)

### *F. Diagnostic Checking*

The next step is to select the best model among all the identified models. For this, residual diagnostics and the model with the maximum log-likelihood and minimum values of Akaike Information Criterion (AIC), modified Akaike Information Criterion (AICc), and Bayesian Information Criterion (BIC) was considered as the best model. Under the residual diagnostics, Ljung-Box Q statistic is used to check whether the residuals are random or not (Tebbs, 2010).

### *G. Akaike Information Criterion (AIC)*

The Akaike's Information Criterion (AIC) says to select the ARIMA (*p*,*d*,*q*) model which minimizes

$$\text{AIC} = -2\ln L + 2k\ldots\ldots\ldots\ldots\ldots\ldots..(6)$$

where ln$L$ is the natural logarithm of the estimated likelihood function and $k = p + q$ is the number of parameters in the model. The AIC is an estimator of the expected Kullback-Leibler divergence, which measures the closeness of a candidate model to the truth. The smaller this divergence, the better the model (Tebbs, 2010).

A problem arises in that AIC is a biased estimator of the expected KL divergence in ARMA (*p*,*d*,*q*) models. An alternative AIC statistic which corrects for this bias is [9],

$$\text{AICC} = \text{AIC} + \frac{2(K+1)(K+2)}{n-k-2}\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(7)$$

### *H. Bayesian Information Criterion (BIC)*

The Bayesian Information Criterion (BIC) says to select the ARIMA (*p*,*d*,*q*) model which minimizes [9],

$$\text{BIC} = -2\ln L + 2k\ln(n)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(8)$$

where ln $L$ is the natural logarithm of the estimated likelihood function and $k = p + q$ is the number of parameters in the model and $n$ is total observations. Both AIC and BIC require the maximization of the log likelihood function and When we compared AICC to BIC offers a stiffer penalty for over parameterized models (Tebbs,2010). An overall check of the model adequacy was made using the modified Box-Pierce ***Q*** statistics. The test statistics is given by:

$$Q_m = n(n+2)\sum_{k=1}^{n}(n-k)^{-1}r_k^2 \approx \chi_{m-r}^2(9)$$

where:

$r_k^2 =$ the residuals autocorrelation at lag $k$
$n =$ the number of residual
$m =$ the number of time lags included in the test.

When the *p*-value associated with the ***Q*** is large the model is considered adequate, else the whole estimation process has to start again in order to get the most adequate model. Here all the tests were performed at the 95% confidence interval (Tebbs, 2010). Furthermore, a plot of the ACF squared residual and PACF squared residuals was performed on the residuals of the fitted model to check for heteroscedasticity and again an ARCH LM-test for conformity of the presence of, or otherwise ARCH effect was performed.

## III. RESULTS AND DISCUSSION

### *A. Stationarity*

A behavioral analysis was carried out on the time series plot and ACF plot and partial PACF plot. The time series plot of Figure 1 shows the trend of road accident cases in Ghana from the period January 1990 to December 2010 and it reveals an existence of non-significant trend. This implies that the data is non-stationary. From the Correlogram (ac) plot of figure 1, it can be seen that the series remains significant for more than half a dozen lags, rather than quickly declining to zero which is also a sign of non-stationarity. Furthermore, the ADF test shown in Table 1 shows that the series was not stationary.

A dummy variable regression equation was used to check for seasonality in the time series data. This was done by regressing road accident cases (dependent variable) against  months of the year (explanatory variables).If at least one of the explanatory variables is significant in the model then there is seasonality in the time series data. Months December and April were found to be significant in the hence the data was seasonal. Figure 1 is a plot of means for the 12 months of the year.  It's clear that there are monthly differences (seasonality).Thus, the data was differenced and tested. As shown in Table 1, the ADF test designates that the data was stationary and deseasonalized after the seasonal and non-seasonal first difference.

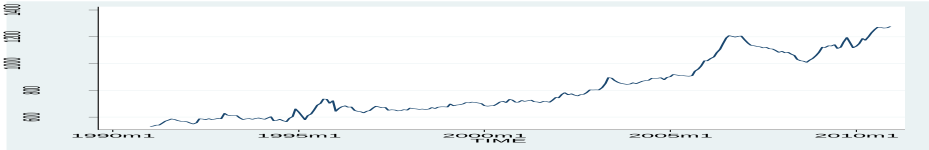


*Figure 1 : Time plot of observed data on road accident cases*

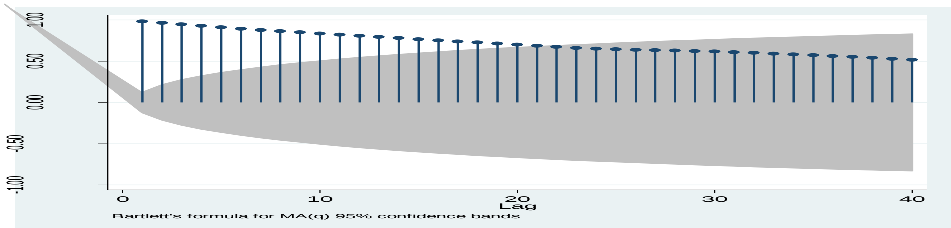


*Figure 2 : Correlogram of accident cases*

*Table 1: Unit Root Tests Results*

| Tests | Order of Difference | Test Statistic | P-Value |
|---|---|---|---|

| | | | |
|---|---|---|---|
| **ADF** | 0 | -3.6853 | 0.9899 |
| **ADF** | 1 | 0.0000 | 0.05293 |

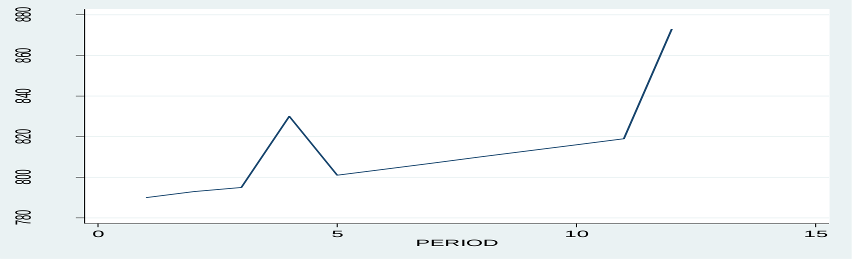


*Figure 3 : Plot of means months*

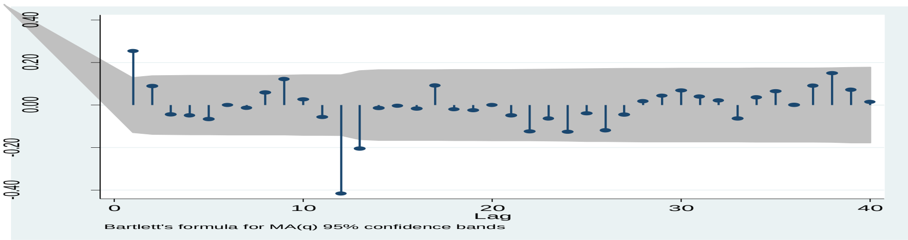


*Figure 4 : Correlogram after first non - seasonal and seasonal differencing*

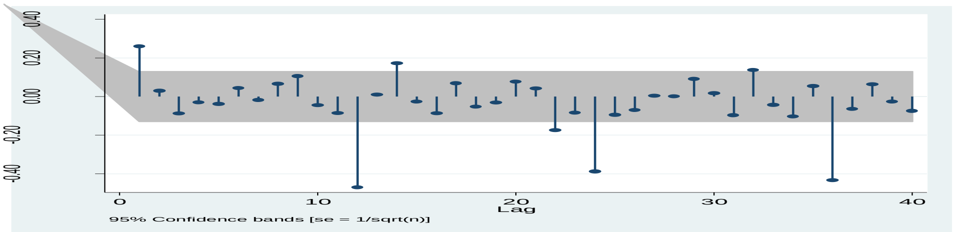


*Figure 5 : Partial Correlogram after first non - seasonal and seasonal differencing*

### *B. Model Identification*

A SARIMA model can be proposed based on the ACF and PACF plot of the differenced data in Figure 5 and 6. Non seasonal behaviour: From both the ACF and PACF graphs show a clear spike at lag .We can guess a non-seasonal AR(1) and MA(1). Seasonal behaviour: From the figures, we are ignoring lags between multiples of 12.We are only interested in lag 12, 24, and 36 in this case. In the ACF, it drops off after one spike, a spike at 12 and nothing at 24. With the PACF there are significant lags at 12, 24, and 36. We can guess this is a seasonal MA(1) .Using this procedure, the models presented in Table 2 were suggested. Taking into consideration the principle of parsimony, different tentative models were proposed.

*Table 2: Different SARIMA Model Fitted*

| Model | AIC | RMSE | MAPE | Log-Likelihood |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| SARIMA(0,1,1)× (0,1,1)12 | 5.898 | 17.980 | 1.687 | 244.32 |
| SARIMA(1,1,1)× (0,1,1)12 | 5.917 | 17.934 | 1.686 | 244.26 |
| **SARIMA(1,1,0)× (0,1,1)12** | **5.892** | **17.930** | **1.688** | **245.48*** |

*Best based on the model selection criterion

From Table 3, **SARIMA (1, 1, 0) × (0, 1, 1)12** was the best model based on the selection criterion used. This is because it satisfies most of the selection criterion. The parameters of this model were then estimated.

*Table 3: Estimated model parameters*

| Model | Component | Coefficient | S.E | Test Statistic | P-Value |
|---|---|---|---|---|---|
| **SARIMA(1,1,0)*(0,1,1)** | constant | 0.0334 | 0.0039 | 8.53 | 0.000 |
| | AR(1) | 0.1946 | 0.0648 | 2.59 | 0.016 |
| | SMA(1) | 0.9044 | 0.0005 | 6.73 | 0.000 |

### *C. Model Diagnostic*

A diagnosis of the model was carried out to see how well it fits the data. From Fig. 6, the ACF of the residuals revealed that the residuals are white noise although a significant spike appeared at lag 0. This could be as a result of random factor. The plot of the Ljung-Box p-values in Fig. 6 furthermore showed that the model was adequate and a good fit for the data. They were above the threshold of 0.05 indicated by the blue line.

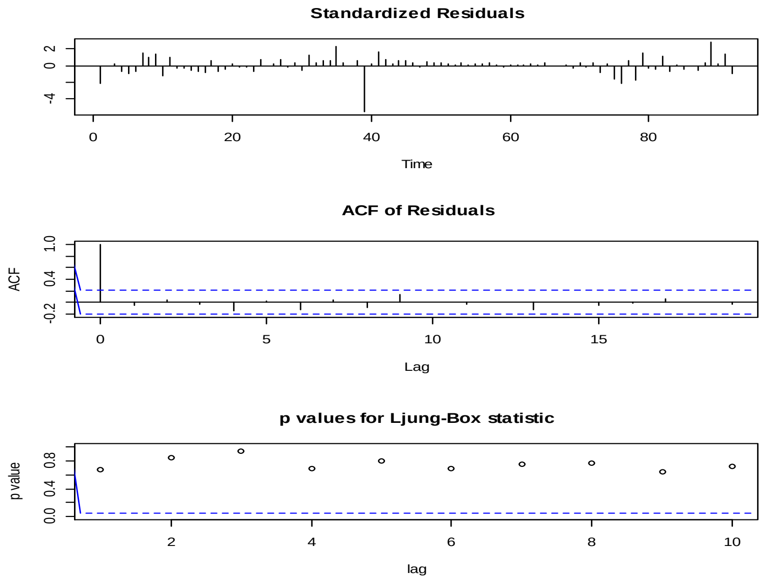


*Figure 6 : Diagnostic plot of residuals of SARIMA (1, 1, 0) × (0, 1, 1)12*

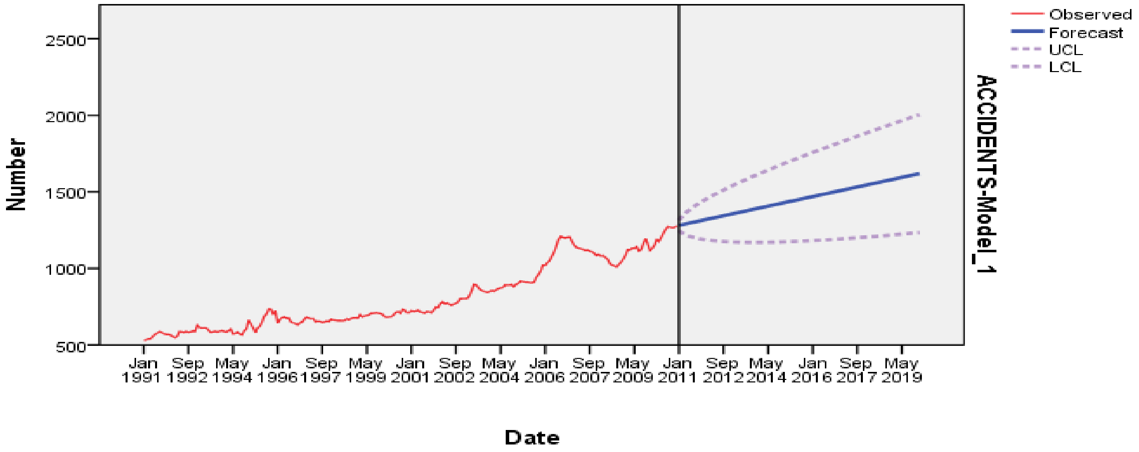
There were no ARCH effects in the ARCH-LM test in Table 4; hence the variance of the residuals are constant. The Ljung-Box p–values (> 0.05) also showed that there a serial correlation does not exist in the residuals of the model. The ACF plot of the residuals in Fig. 6 also shows that the residuals are white noise series.

*Table 4: ARCH LM Test for of residuals*

| Model | Chi-squared | P-Value |
|---|---|---|
| **SARIMA(1,1,0)× (0,1,1)$_{12}$** | 20.7525 | 0.9997 |

### *D. Forecasting*

The model was fitted for a ten year period after the diagnostic test. Below is the graph which gives a pictorial view of the observed series, its forecast, the confidence intervals of the forecast. It is predicted that the incidence road accidents will gnerally increase in Ghana for the period forecasted.



## IV. CONCLUSION

Based on the results of this study, the rate of road accidents is expected to increase at least for the next years 10 years. It was found that the incidence of road accidents and injuries as result of accidents in Ghana can be fitted to a SARIMA (1, 1, 0)*(0, 1, 1)$_{12}$ model. From the findings it can be deduced that the carnage on our roads will increase and this will have serious repercussion on the country. It is expected that all stakeholders will find the study useful.

### References

1. Abbas, K. A. (2004). Traffic safety assessment and development of predictive models for accidents on rural roads in Egypt. *Accident Analysis & Prevention*, *36*(2), 149-163.
2. Adu-Poku, K. A., Avuglah, R. K., & Harris, E. (2014) Modeling Road Traffic Fatality Cases in Ghana.
3. Agyemang, B., Abledu, G. K., & Semevoh, R.(2013) Linking Road Traffic Accidents to Vehicle Population; Empirical Evidence from Ghana.

4. Box, G. E., & Jenkins, G. M. (1976). *Time series analysis: forecasting and control, revised ed*. Holden-Day.
5. Fadugba, O. M. (2012). Time Series Analysis of the Rate of Road Traffic Accident in Nigeria (A Case Study of Lagos Stale).
6. Ho, S. L., Xie, M., & Goh, T. N. (2002). A comparative study of neural network and Box-Jenkins ARIMA modeling in time series prediction. Computers & Industrial Engineering, 42(2), 371-375
7. Kehinde Ibukun Olobatuyi(2012), Federal University of Agriculture Abeokuta (FUNAAB), Ogun State, Nigeria
8. Mahadeva L. and Robinson P. (2014), Unit root testing to help model building (centre for central Bank Studies, Bank of England, 2004).
9. Mustafa, M. N. (2005). Overview of current road safety situation in Malaysia. *Highway planning Unit, Road Safety Section, Ministry of Works*, 5-9.
10. Okutu, J. K. (2011). *Time series analysis of road traffic accidents in ghana, a case study of accra–tema motorway, greater accra region thesis submitted to the institute of distance learning* (Doctoral dissertation, INSTITUTE OF DISTANCE LEARNING, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY).
11. Peden, M. (2004). *World report on road traffic injury prevention*.
12. Tebbs M. (2010).STAT 520 Forecasting and Time Series (University of South Carolina, Department of Statistics, spring , 2010).