Application of ARIMA Model on Monthly Rainfall Data

I. O. Shittu[1]; A. O. Bello[2]; M. A. Obomeghie[3]; G. A. Yinusa[4]

[1]Department of Statistics,
Yaba College of Technology, Yaba, Lagos, Nigeria.
e-mail: idoscomus2011@gmail.com

[2]Department of Statistics,
University of Ibadan,
Ibadan, Nigeria.

Abstract **— This paper aimed at modelling monthly rainfall of three locations in Nigeria: Asaba, Benin and Port-Harcourt. Time series model; Auto-Regressive Integrated Moving Average (ARIMA) model was adopted. Monthly rainfall data generated from statistical database of Central Bank of Nigeria in some towns were used. The trend rainfall in Port-Harcourt, Asaba and Benin exhibit Non-stationarity. First difference was taken for all the locations to achieve stationarity. The Autocorrelation Function ACF plot, Partial Autocorrelation Function PACF, from the model selection criterion ARIMA (1,1,3) model is best use to fit a rain fall data for duration 2007 – 2012. In Port-Harcourt, Asaba and Benin region an ARIMA (1,1,2) model is best use in fitting the rainfall data. The output of model fitted shows that Asaba estimate of AIC, BIC, AICC, loglikelihood and sigma square perform better than the Benin with the same ARIMA model of (1,1,3). The residual plot of ACF and PACF shows that the data are normally, independent and identically distributed. The predicted forecast of figure 8, 15 and 21 shows a seasonal trend of future value which is present in the real data from the rainfall.**

Keywords - *Rainfall, Location, ARIMA Model.*

I. Introduction

Dry seasons and raining seasons are distinguished period of the year (January Pto December). Generally, in some parts of Nigeria the 5 months period (October, November, December, January and February) are considered as dry season while 7 months period of the year (March, April, May, June, July, August and September) are taken as rainfall season depending on the weather and climatic situation.

Climatic change do bring about variation in the length and volume of rainfall, creating shorter or longer rainfall periods. This variation among other factors affect agricultural production, weather condition in terms of temperature, humidity. The geographical location of a place also affect the length of seasonal durations. This paper compare three locations (towns) in Nigeria; Asaba, Benin and Port-Harcourt monthly rainfall.

II. Data Source and Data Manipulation

Secondary sources of data collection was adopted. The data on Rainfall in three locations in some Nigeria Towns were obtained from the data base of the Central Bank of Nigeria through their webpage at http://www.cbn.gov.ng/search/runsearch.asp?q=rainfall%20data.

These dataset was collected for the periods spanning 2007 to 2014 on the amount of monthly rainfalls in Nigeria.

III. Materials and Methods

A time series is a sequence of ordered data. The "ordering" refers generally to time. We made use of Time Series Analysis to detect patterns of change in statistical information over the regular interval of time. We project the pattern to arrive at an estimate for the future. All statistical forecasting methods are extrapolatory in nature i.e they involve the projection of past patterns or relationship into the future. Time series data can be stationary and non-stationary. However, theory of time series is concerned with stationary time series. A time series data $\{Y_t\}$ is said to be

stationary if there is no systematic change in its mean and variance and if all periodic variations have been removed. Such series is assumed to have been in a state of statistical equilibrium where the statistical properties of a stationary process does not change over time Shittu (2011). A time series is said to be stationary if it has constant mean and variance (Osabuohien, 2013); the paper further elaborated that a stationary time series $\{Y_t\}$ follows an autoregressive moving average model of order p and q (denoted as ARIMA (p, d, q) if it satisfies the difference equation;

$$Y_t = \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \phi_3 Y_{t-3} - \ldots - \phi_p Y_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \quad (1)$$

$$Y_t - \sum_{i=1}^{p} \phi_i Y_{t-i} = \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \quad (2)$$

$$\phi(B) Y_t = \theta(B) \varepsilon_t \quad (3)$$

Y is an ARIMA (p,d,q) process if $\nabla^d Y_t$ is ARIMA (p, d, q)

$$\phi(B) \nabla^d Y_t = \theta(B) \varepsilon_t \quad (4)$$

$$\phi(B)(1-B)^d Y_t = \theta(B) \varepsilon_t \quad (5)$$

Where ϕ_i and θ_j are constants such that the zeros of equation are all outside the unit circle for stationarity and invertibility respectively. For a seasonal series, the time plot reveals the existence of a seasonal nature in data and the ACF or Correlogram exhibits a spike at the seasonal lag. Box (1976), Madsen, (2008), Meese (1982) have contributed in formulation of the theory and practice of Time Series Analysis (TSA) of ARIMA models. The knowledge of the theoretical properties of the models provides basis for their identification and estimation (Osabuohien, 2013).

Auto-Regressive Process

Autoregressive processes are regressions on themselves. A p-order autoregressive process AR(p). {Yt} satisfies the equation (Yule (1926))

$$Y_t = \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \Phi_3 Y_{t-3} + \ldots + \Phi_p Y_{t-p} + e_t \quad (6)$$

The current value of the series Y_t is a linear combination of the p most recent past values of itself plus an "innovation" term that incorporates everything new in the series at time t that is not explained by the past values. Thus, for every t, we assume that is independent of Y_{t-1}, Y_{t-2}, Y_{t-3}, ..., Y_{t-q}.

Moving Average Processes

Moving average models were first considered by Slutsky (1927) and Wold (1938) as cited by Mohammed (2014). The Moving Average Series can be written as

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \theta_3 e_{t-3} - \ldots - \theta_q e_{t-q} \quad (7)$$

We call such a series a moving average of order **q** and abbreviate the name to **MA(q)**. where, Y_t is the original series and et is the series of errors. The current value of the series Y_t is a linear combination of the p most recent past values of itself plus an "innovation" term that incorporates everything new in the series at time t that is not explained by the past values. Thus, for every t, we assume that is independent of Y_{t-1}, Y_{t-2}, Y_{t-3}, ..., Y_{t-q} .

Autoregressive Integrated Moving Average (ARIMA) Model

The Box and Jenkins (1970) procedure is the milestone of the modern approach to time series analysis. Given an observed time series, the aim of the Box and Jenkins procedure is to build an ARIMA model. In particular, passing by opportune preliminary transformations of the data, the procedure focuses on Stationary processes. We fitted the Box-Jenkins Autoregressive Integrated Moving Average (ARIMA) model. This model is the generalized model of the non-stationary ARMA model denoted by ARMA (p,q) can be written as

$$Y_t = \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \Phi_3 Y_{t-3} + \ldots + \Phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \theta_3 e_{t-3} - \ldots - \theta_q e_{t-q} \quad (8)$$

Where, Yt is the original series, for every t, we assume that is independent of Y_{t-1}, Y_{t-2}, Y_{t-3}, …, Y_{t-p} .

A time series $\{Y_t\}$ is said to follow an integrated autoregressive moving average (ARIMA) model if the dth difference $W_t = \nabla d Y_t$ is a stationary ARMA process. If $\{W_t\}$ follows an ARMA (p,q) model, we say that $\{Y_t\}$ is an ARIMA(p,d,q) process. Fortunately, for practical purposes, we can usually take d = 1 or at most 2.

Consider then an ARIMA (p,1,q) process. With $W_t = Y_t - Y_{t-1}$ we have

$$W_t = \Phi_1 W_{t-1} + \Phi_2 W_{t-2} + \Phi_3 W_{t-3} + \ldots + \Phi_p W_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \theta_3 e_{t-3} - \ldots - \theta_q e_{t-q} \quad (9)$$

Box and Jenkins procedure's steps

i. *Preliminary analysis:* create conditions such that the data at hand can be considered as the realization of a stationary stochastic process.

ii. *Identification:* specify the orders p, d, q of the ARIMA model so that it is clear the number of parameters to estimate. Recognizing the behavior of empirical autocorrelation functions plays an extremely important role.

iii. *Estimate:* efficient, consistent, sufficient estimate of the parameters of the ARIMA model (maximum likelihood estimator).

iv. *Diagnostics:* check if the model is a good one using tests on the parameters and residuals of the model. Note that also when the model is rejected, still this is a very useful step to obtain information to improve the model.
v. *Usage of the model:* if the model passes the diagnostics step, then it can be used to interpret a phenomenon, forecast.

Residuals Diagnostic Checking

Jarque-Bera Test

We can check the normality assumption using Jarque-Bera (Jarque & Bera, 1980) test which is a goodness of fit measure of departure from normality, based on the sample kurtosis (k) and skewness(s). The test statisticsJarque-Bera (JB) is defined as

$$JB = \frac{n}{6}\left(s^2 + \frac{(k-3)^2}{4}\right) \sim \chi^2_{(2)} \qquad (11)$$

Where n is the number of observations and k is the number of estimated parameters. The statistic JB has an asymptotic chi-square distribution with 2 degrees of freedom, and can be used to test the hypothesis of skewness being zero and excess kurtosis being zero, since sample from a normal distribution have expected skewness of zero and expected excess kurtosis of zero.

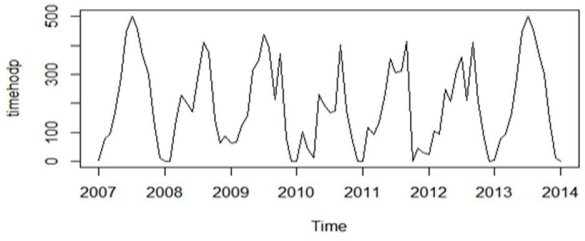
Ljung-Box Test

Ljung-Box Test can be used to check autocorrelation among the residuals. If a model fit well, the residuals should not be correlated and the correlation should be small Box and Ljung, 1978. In this case the null hypothesis is
H0 : $\rho_1(e) = \rho_2(e) = \ldots\ldots = \rho_k(e) = 0$ is tested with the Box-Ljung statistic

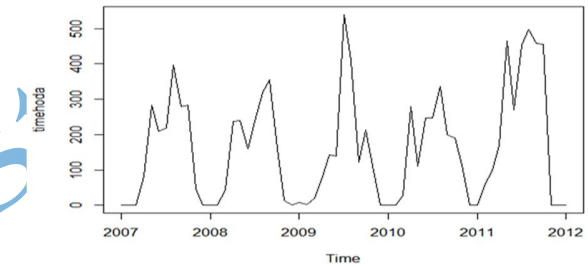
$$Q = N(N+1)\sum_{i=1}^{k}(N-k)\rho_k^2(e) \qquad (12)$$

Where, N is the no of observation used to estimate the model. This statistic Q approximately follows the chi-square distribution with (k-q) df, where q is the no of parameter should be estimated in the model. If Q is large (significantly large from zero), it is said that the residuals autocorrelation are as a set are significantly different from zero and random shocks of estimated model are probably auto-correlated. So one should then consider reformulating the model.

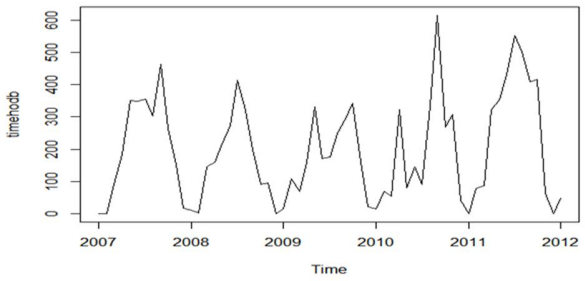
IV. ANALYSIS, RESULTS AND DISCUSSION



a. Port-Harcourt



b. Asaba



c. Benin City

Figure 1 : Time Series Plot of Monthly Rainfall 2007 to 2014 of three locations

The data series plots indicate non-stationary in the three locations; as the series wanders up and down for long periods. Consequently, we will take a first difference of the data. The differenced data plot is shown in Figure 2.

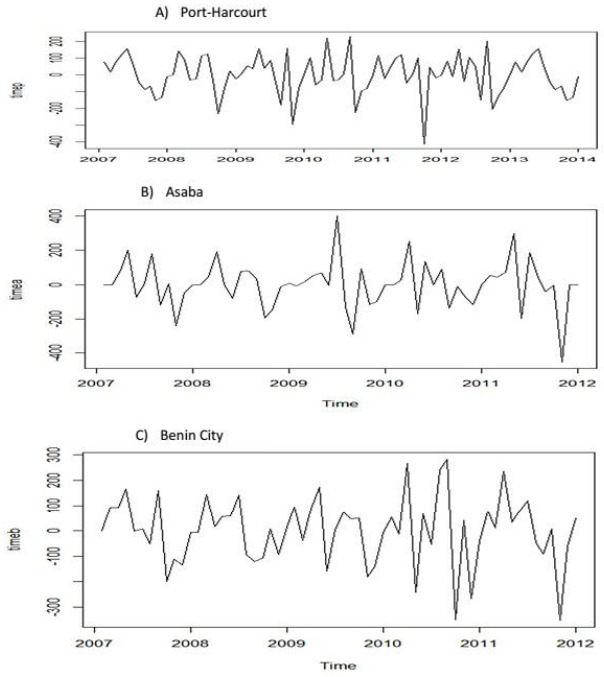


Figure 2: First Difference Plot of Monthly Rainfall 2007 to 2014 of the locations.

The first difference of the series shows some level of stationarity in the data in the three locations.

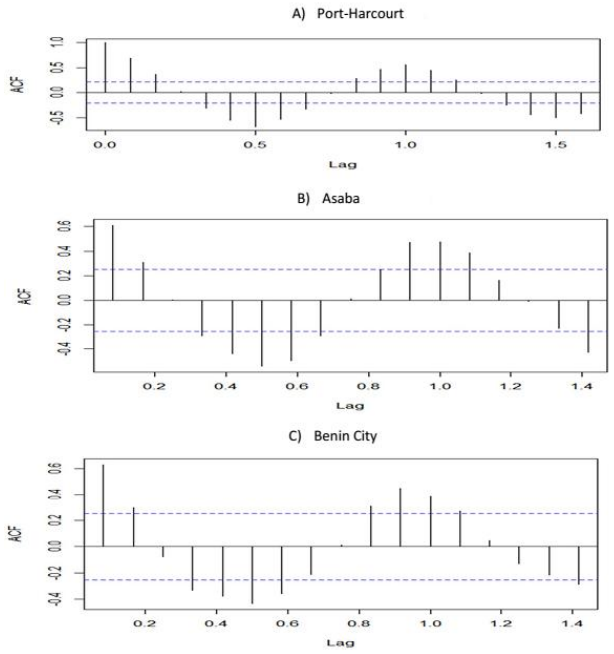


Figure 3: ACF Plot of Monthly Rainfall 2007 to 2014 of the locations

The ACF plots in Fig 3 indicate Moving average of order 3 i.e. MA (3)

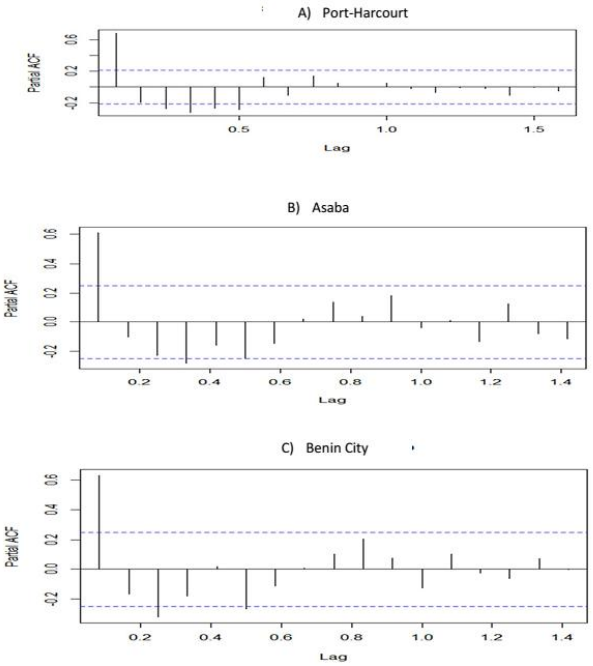


Figure 4: PACF Plot of Monthly Rainfall 2007 to 2014 of the locations.

Augmented Dickey-Fuller Test

Dickey-Fuller = -6.8576, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary

The test above shows a stationary series.

The ACF and PACF shown in Figure 3 and 4 is suggestive of an AR (1) and MA (3) model. So an initial candidate model is an ARIMA (1,1,3). There are no other obvious candidate models.
The PACF and ACF shown in Figure 3 and 4 is suggestive of an AR(1) and MA (2) model. So an initial candidate model is an ARIMA (1,1,2). There are no other obvious candidate models.

Table 1: Table of Coefficients

Locations	AR1	MA1	MA2	MA3	SMA1	Sigma Sqr	Log likelihood	AIC	BIC
Port Harcourt	-0.7158	-1.0905	-0.8006	0.8915	-0.9996	5198	-424.3	858.6	872.17
Asaba	0.1252	-1.9937	1.0000		-0.9991	6923	-289.96	587.93	597.18
Benin City	0.3492	-1.9850	1.0000		-0.9999	9651	-296.18	600.35	609.61

V. CONCLUSION

The Time plot in Figure 1, 9 and 16 shows the original trend (pattern) of rainfall data in Port-Harcourt, Asaba and Benin exhibit Non-stationarity. First difference was taken for all the locations to make it Stationary, that was shown in Figure 2, 10 and 17.

The Autocorrelation Function ACF plot in Figure 3, 11 and 17. Partial Autocorrelation Function PACF Figure 4, 12 and 18. From the model selection criterion ARIMA(1,1,3) model is best use to fit a rain fall data for duration 2007 – 2012. In Port-Harcourt, Asaba and Benin region an ARIMA (1,1,2) model is best use in fitting the rainfall data.

The output of model fitted shows that Asaba estimate of AIC, BIC, AICC, loglikelihood and sigma square perform better than the Benin with the same ARIMA model of (1,1,3). The residual plot of ACF and PACF shows that the data are normally, independent and identically distributed. The predicted forecast of figure 8, 15 and 21 shows a seasonal trend of future value which is present in the real data from the rainfall.

REFERENCES

Box, G. E, P, and Jenkins, G. M. (1976): Time Series Analysis, Forecasting and Control,. Holden-Day, San Francisco.

Box, G. E. P., & Jenkins, G. M. (1976), Time Series Analysis, Forecasting and Control, San

Forecasting Power and Risk Premiums *JEL Classifications, University of Valencia*.

Francisco, Holden-Day, California, USA

Hamjah, M.A. and Chowdhury, M.A.K.(2014), Measuring Climatic and Hydrological Effects on Cash Crop Production and Production Forecasting in Bangladesh Using ARIMAX Model,

Jarque, Carlos M. Bera, Anil K. (1980), Efficient Test for Normality, Homoscedasticity and Serial Independence of Regression Residuals, Economics Letters, 6(3); 255 – 259.

Julio J. Lucia and Hipolit Torro (2005), Short Term Electricity Future Prices at Nord Pool: *Mathematical Theory and Modeling,* 4(6):138-152.

Madsen H (2008), Time Series Analysis, Chapman & Hall/CRC, London.

Meese R. and J. Geweke (1982), A Comparison of Autoregressive Univaraite Forecasting Procedures for Macroeconomic Time Series. University of California, Berkeley, CA, USA.

Mohammed A. H. (2014), Temperature and Rainfall Effects on Spice Crops Production and Forecasting the Production in Bangladesh: An Application of Box-Jenkins ARIMAX Model Mathematical Theory and Modeling www.iiste.org ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online) Vol.4, No.10, 2014

Osabuohien-Irabor O. (2013), Applicability of Box Jenkins SARIMA Model in Rainfall Forecasting: A Case Study of Port-Harcourt South South Nigeria, Canadian Journal on Computing in Mathematics, Natural Sciences, Engineering and Medicine Vol. 4 No. 1, February 2013.

Shittu I. O. & Yaya S. O. (2011), Introduction to Time Series Analysis. Fasco Printing Works, 67 Gbadebo Street, Mokola, Ibadan.

Yule, G, U, (1926), Why do we sometimes get nonsense-correlations between time-series? A study in sampling and the nature of time-series. Journal of the Royal Statistical Society, 89