International Journal of Mathematics and Physical Sciences Research ISSN 2348-5736 (Online)
Vol. 6, Issue 1, pp: (106-112), Month: April - September 2018, Available at: www.researchpublish.com

# A STUDY OF THE SUITABLE TIME SERIES MODEL FOR MONTHLY CRUDE OIL PRODUCTION IN NIGERIA

Sadeeq S. A. and Ahmadu A. O.

Email Address: sadeeqsa@feca.edu.ng

***Abstract:*** **Petroleum production and export play a dominant role in Nigeria's economy and account for about 90% of her gross earnings. This dominant role has pushed agriculture, the traditional mainstay of the economy, from the early fifties and sixties, to the background. In this research work we fitted a univariate Seasonal Autoregressive Integrated Moving Average model (SARIMA) to the monthly crude oil production in Nigeria between 2002 and 2016. Different Box-Jenskin Autoregressive Integrated Moving Average (ARIMA) models are fitted and diagnosed. However, ARIMA $(2,1,0)(2,1,1)_{12}$ was the best model for the data. The model was further validated and it was discovered that autocorrelation between residuals at different lag times was not significant. Finally, the time plot of the in-sample forecast errors shows that the variance of the forecast errors seems to be roughly constant over time and the histogram of the time series shows that the forecast errors are roughly normally distributed and the mean seems to be close to zero, it seems plausible that the forecast errors are normally distributed with mean zero and constant variance.**

***Keywords:*** **SARIMA, petroleum, Box-jenskin, model, residuals, autocorrelatiom.**

## 1. INTRODUCTION

Modern time series forecasting methods are essentially rooted in the idea that the past tells us something about the future. The question of how exactly we are to go about interpreting the information encoded in past events, and furthermore, how we are to extrapolate future events based on this information, constitute the main subject matter of time series analysis. Typically, the approach to forecasting time series is to first specify a model, although this need not be so. This model is a statistical formulation of the dynamic relationships between that which we observe (i.e. the so called information set), and those variables we believe are related to that which we observe. It should thus be stated immediately that this discussion will be restricted in scope to those models which can formulated parametrically. The "classical" approach to time series forecasting derives from regression analysis. The standard regression model involves specifying a linear parametric relationship between a set of explanatory variables and the dependent variable. The parameters of the model can be estimated in a variety of ways, going back as far as Gauss in 1794 with the "Least Squares" method, but the approach always culminates in striving for some form of statistical orthogonality between the explanatory variables and the residuals (or innovations) of the regression. That is, we wish to express the linear relationship in a dichotomous form in which the innovations represent that part of our information which is completely unpredictable. It should probably also be emphasized that in the engineering context this is analogous to reducing a signal to "white noise." However, this review is to be concerned with more "modern" approaches and in many ways, it was the practical necessities of engineering that provided an initial impetus. Both Wiener (1949) and Kolmogorov (1941) were pioneers in the field of linear prediction, and while their approaches differed (Wiener worked in the frequency domain popular amongst engineers, while Kolmogorov worked in the time domain), it is clear that their solutions to the same basic geometrical problem were equivalent (see Priestley (1981) ch.10). Wiener's work, in particular, was especially relevant to modern time series forecasting in that he was among the first to rigorously formulate the problem of "signal extraction." That is, given observations on a time series corrupted by additive noise, what is the optimal estimator (in the Mean-Squared Error (MSE) sense) of the latent or underlying signal (or state variable). Given the historical context of massive systems of

equations models popular among macro econometric forecasters of 1950's (see for example the Klein-Goldberger model (1955) or Adelmans (1959) for details) it quickly became apparent that forecasting models derived from a signal extraction context forecasted at least as well as those based on complicated systems of economic relationship equations formulated as individual, yet interconnected, dynamic classical regressions.

Following the discovery of crude oil by Shell D'Arcy Petroleum, pioneer production began in 1958 from the company's oil field in Oloibiri in the Eastern Niger Delta. By the late sixties and early seventies, Nigeria had attained a production level of over 2 million barrels of crude oil a day. Although production figures dropped in the eighties due to economic slump, 2004 saw a total rejuvenation of oil production to a record level of 2.5 million barrels per day. The Monthly Petroleum Information (MPI) and the Annual Statistical Bulletin (ASB) represent the activities of the Oil and Gas Industry in Nigeria. These in essence give a clear picture of the activities that spell out the major economic profile of Nigeria driving for transparency and accountability; to give would-be and/or prospective investors a bird's–eye-view of possible investment opportunities in the Country. It includes general review of the Petroleum Industry activities, data on seismic activities, crude oil production, liftings, allocations, exports by destination, receipts, Gas production, utilization, sales, transmission and exports. The main objective of this research work is to determine the best suitable model for the monthly crude oil production in Nigeria. And to investigate this, different Box-Jenskin Autoregressive Integrated Moving Average (ARIMA) models are fitted and diagnosed respectively.

## 2. METHODOLOGY

The model used in this study is the ARIMA proposed by Box and Jenkins (1976). The preliminary test for stationarity and seasonality of the data was conducted in which differences (d) as well as seasonal differences(D) were taken. After the stationarity of the series was attained, ACF and PACF of the stationary series are employed to select the order p and q of the ARIMA model. At this stage, different candidates' model manifested and their parameters are estimated using the maximum likelihood method. Based on the principle of parsimony and model diagnostic tests, we obtained the best fitting ARIMA model.

**Notation for ARIMA Models**

A dependent time series that is modeled as a linear combination of its own past values and past values of an error series is known as a (pure) ARIMA model.

**Non-seasonal ARIMA Model Notation**

The order of an ARIMA model is usually denoted by the notation ARIMA (p ,d, q), where p is the order of the autoregressive part

d is the order of the differencing (rarely should d > 2 be needed)

q is the order of the moving-average process

Given a dependent time series { $Y_t : 1 \leq t \geq n$ } ,mathematically the ARIMA model is written as

$$(1 - B)^d Y_t = \mu + \frac{\theta(B)}{\emptyset(B)}\alpha_t$$

Where:

t is indexes time

μ is the mean term

B is the backshift operator; that is, $BX_t = X_{t-1}$

$\emptyset(B)$ is the autoregressive operator, represented as a polynomial in the back shift operator:

$\emptyset(B) = 1 - \emptyset_1(B) - \ldots - \emptyset_p(B)^p$

$\theta(B)$ is the moving-average operator, represented as a polynomial in the back shift operator:

$\theta(B) = 1 - \theta_1(B) - \ldots - \theta_q(B)^q$

$a_t$ is the independent disturbance, also called the random error

**Seasonal ARIMA Model Notation**

Seasonal ARIMA models are expressed in factored form by the notation ARIMA (p,d,q)(P,D,Q)$_s$, where

P is the order of the seasonal autoregressive part

D is the order of the seasonal differencing (rarely should D > 1 be needed)

Q is the order of the seasonal moving-average process

s is the length of the seasonal cycle

Given a dependent time series $Y_t : 1 \leq t \geq n$ },mathematically the ARIMA seasonal model is written as

$$(1 - B)^d(1 - B)^D Y_t = \mu + \frac{\theta(B)\theta_s(B^s)}{\emptyset(B)\phi_s(B^s)}\alpha_t$$

Where:

$\phi_s(B^s)$ is the seasonal autoregressive operator, represented as a polynomial in the back shift operator:

$\phi_s(B^s) = 1 - \phi_{s,1}B^s - \cdots - \phi_{s,p}B^{s\,p}$

$\theta_s(B^s)$ is the seasonal moving-average operator, represented as a polynomial in the back shift operator:

$\theta_s(B^s) = 1 - \theta_{s,1}\,B^s - \cdots - \theta_{s,p}\,B^{sp}$

## 3. MODEL SELECTION CRITARIA

In time series analysis there may be several adequate models which can be used to represent a given data set. Sometimes the best choice is easy, other times the best choice can be very difficult. Thus various criterions for model assumption have been introduced in the literature for model selection. They are different from the model identification methods, model identification tools such as ACF, PACF, are used only for identifying adequate models. Residuals from all adequate models are white noise for a given data set. When there are multiple adequate models, the selection criterion is normally based on summary statistics from residuals computed from a fitted model or on forecast errors calculated from out- sample forecast.

**Akaike's Information Criterion (AIC)**

Assume that statistical model of m parameters is fitted to the data. To assess the quality of model fitting, Akaike (1973, 1974a) introduced an information criterion. The criterion has been called (AIC) Akaike's information criterion in the literature it is defined as

AIC (K) = -2ln{maximum likelihood} + 2K

Where K is the number of parameters in the model. Minimizing Ln(L), AIC criterion reduce to

AIC(K)= $n\text{Ln}\hat{\sigma}_n^{\,2}$ + 2K

The optimal order of the model is chosen by the value of K, which is a function of p and q so that AIC(K) is minimum, where L stands for likelihood function of sample and n is the effective number of observations.

## 4. DIAGNOSTIC CHECKING

Time Series modeling is an iterative procedure. To start model identification and parameter estimation, after estimation of parameters we have to assess model adequacy by detecting whether the model assumptions are satisfied. The basic assumption is that $\{z_t\}$ are white noise, that is $\{z_t\}$ are uncorrelated random shock with zero mean and constant variance. For any estimated model, the residual $z_t$'s are estimates of uncorrelated white noise $z_t$'s. Hence model diagnostic checking is accomplished through careful analysis of residual series $\{z_t\}$. Because this residual series is the product of parameter estimation. To check whether the errors are normally distributed, one can construct a histogram of standardized residual $\frac{\hat{z}_t}{\hat{\sigma}_z}$ and compare with the standard normal distribution using the chi-squared goodness of fit test. To check whether the variance is constant, we can examine the plot of residuals, to check whether residuals are white noise, we can compute

sample ACF and sample PACF of the residuals to see whether they do not form any pattern and all are statistically insignificant within confidence limits. The sample ACFs, $\hat{\rho}_k$, of residual of required model, lie within these limits

$$- Z_{\frac{\alpha}{2}} \frac{1}{\sqrt{T}} \leq \hat{\rho}_k \geq Z_{\frac{\alpha}{2}} \frac{1}{\sqrt{T}}$$

Another useful test is the Portmanteau lack of fit test. This test uses all the residual sample ACF's a zero to check the Null Hypothesis:

$H_0 = \rho_1 = \rho_2 = \rho_3 = \ldots. = \rho_k = 0$:

$H_1$ not $H_0$

With test statistic $Q = n(n+2)\sum_{j=i}^{k}(n-j)^{-1}\hat{\rho}_j^{\,2}$

This test statistic is the modified Q statistic originally proposed by Box and Pierce (1970). Under the Null hypothesis of model adequacy, Ljung and Box (1978) and Ansely and Newbold (1974) show that Q statistic approximately follows the $\chi^2_{K-M}$ chi- square distribution based on K-M degrees of freedom, where M is the number of parameters estimated in the model

$Q \approx \chi^2_{K-M}$

$H_0$ reject if Q>K with α level of significance, where $K=\chi^2_{\alpha}$,K-M based on the results of these residual analysis, if the present model can be derived.

## 5. RESULT AND DISCUSSION

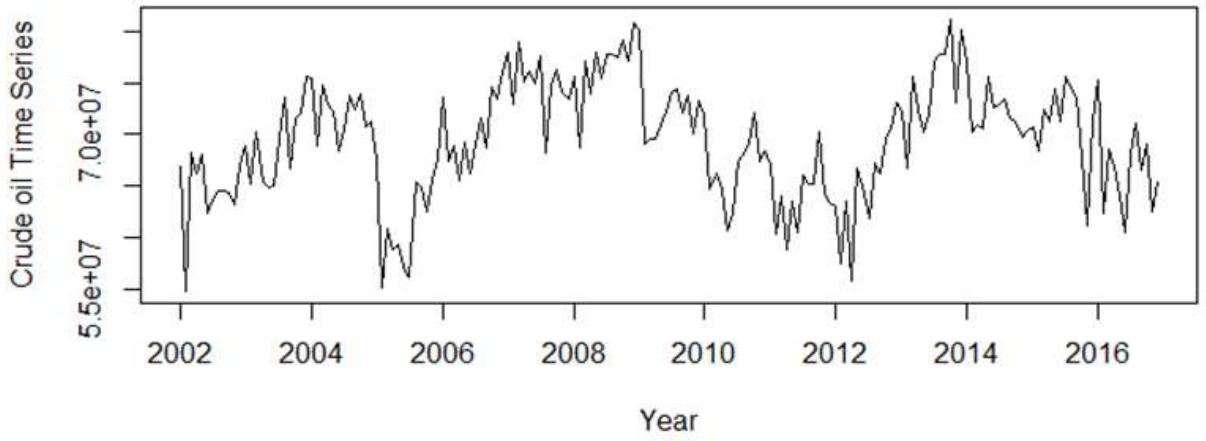



**FIGURE 1.0: TIME PLOT OF MONTHLY CRUDE OIL PRODUCTION IN NIGERIA**

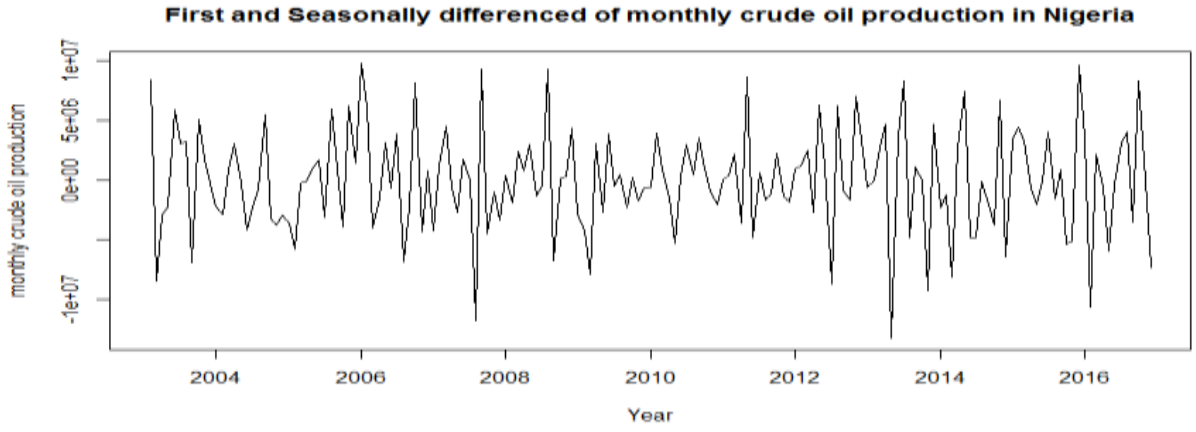



**FIGURE 2.0: TIME PLOT OF FIRST AND SEASONAL DIFFERENCED OF MONTHLY CRUDE OIL PRODUCTION IN NIGERIA**

**STATIONARITY**

The resulting time series of the crude oil production in the figure 1 does not appear to be stationary in mean. Augmented Dickey-Fuller Test shows a p-value= 0.01.Therefore, we can difference the time series once, to see if that gives us a stationary time series. The time plot of first and seasonal differences from the figure 2 above does appear to be stationary in mean and variance, as the level of the series stays roughly constant over time, and the variance of the series appears roughly constant over time. Thus, it appears that we need to difference the time series of the crude oil production once in order to achieve a stationary series. So a seasonal ARIMA (p, 1, q) (P, 1, Q) [12] model is probably appropriate for the time series of the monthly crude oil production in Nigeria.

**CORRELOGRAM**

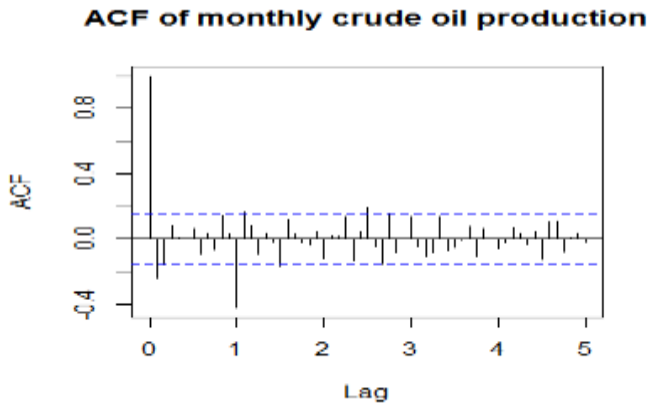



**FIGURE 3.0: ACF OF MONTHLY CRUDE OIL PRODUCTION**

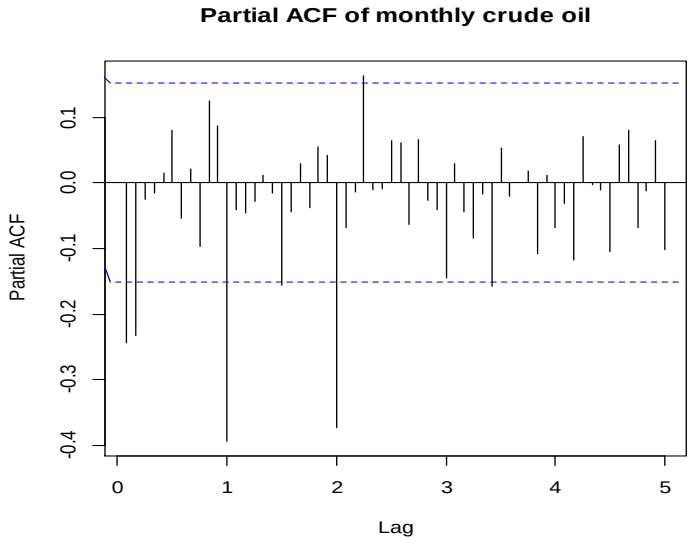



**FIGURE 4.0: PARTIAL ACF OF MONTHLY CRUDE OIL PRODUCTION**

**MODEL IDENTIFICATION**

From figure 3.0 & 4.0, The ACF and PACF of the differenced data follow. The interpretation:

- Non-seasonal: Looking at just the first 2 or 3 lags, it seems possible that an AR(2) might work based on the two spikes in the PACF. On the other hand, the large spike in the ACF at lag 1 might lead to an MA(1) interpretation.

- Seasonal: Look at lags that are multiples of 12 (we have monthly data). Not much is going on there, although there is a significant spike in the ACF at lag 12. Nothing significant is happening at the higher lags. Maybe a seasonal MA(1) might work. On the other hand, the first two seasonal lags in the PACF might suggest a seasonal AR(2). we may fit an ARIMA (2, 1, 1)(2, 1, 1)[12] to the original time series.

**OVERFITTING**

The overfitting process is a technique used in time series to test for other possible significant terms (that may have been missed at the model identification stage of the analysis). We begin by overfitting an MA term on the model (i.e. we fit an ARIMA (2,1,1)(2,1,2)[12] to the original time series). If this additional MA term is not significantly different from zero, we therefore revert our original data and overfit AR. (i.e. we overfit an ARIMA(2,1,0)(2,1,1)[12] to the original time series).

**TABLE 1.0**

| MODEL | AICc |
|---|---|
| ARIMA(2,1,1)(2,1,1)$_{12}$ | -525.87 |
| ARIMA(2,1,2)(2,1,1)$_{12}$ | -523.67 |
| ARIMA(2,1,0)(2,1,1)$_{12}$ | -527.97 |
| ARIMA(2,1,1)(1,1,1)$_{12}$ | -525.58 |
| ARIMA(2,1,1)(1,1,0)$_{12}$ | -482.8 |
| ARIMA(2,1,0)(2,1,0)$_{12}$ | -512.67 |

**FINAL MODEL**

We may therefore conclude that ARIMA (2,1,0)(2,1,1)$_{12}$ is the best ARIMA model so far (i.e because it has the smallest AICc value in table 1.0). The model ARIMA(2,1,0)(2,1,1)12 is of the following form:

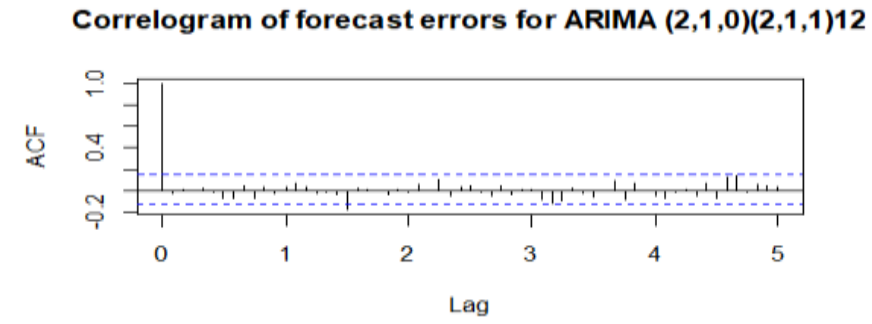
$(1- \Phi_1 B^{12}-\Phi_2 B^{24})(1- \emptyset_1 B-\emptyset_2 B^2)(1-B^{12})(1\text{-B})\, \boldsymbol{Y_t}= (1- \Theta_1 B^{12})a_t$

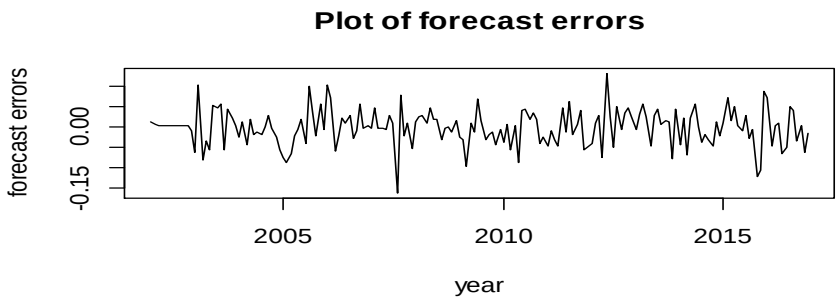
**TABLE 2.0**

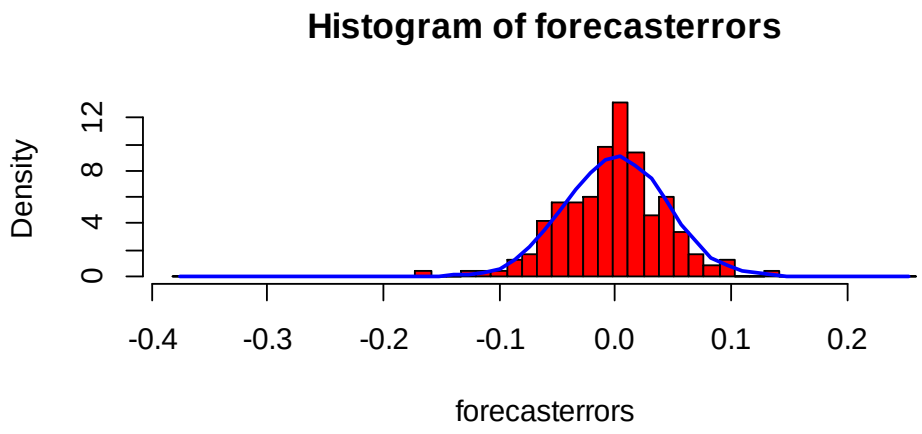
| Seasonal ARIMA Structure | Parameters Estimate | Sigma^2 | log likelihood | Standard Error of Estimate | AICc |
|---|---|---|---|---|---|
| (2,1,0)(2,1,1)[12] | $\Theta_1$=- 0.8463 | 0.002049 | 270.25 | 0.0761 0.0789 | -527.97 |
| | $\phi_1$= − 0.2719, $\phi_2$= - 0.2290<br>$\Phi_1$= - 0.0413, $\Phi_2$= - 0.1689 | | | 0.1247<br>0.1015 0.1247 | |

**MODEL DIAGNOSTICS**

Residual diagnostic tests are used here to determine the goodness-of-fit of the ARIMA (2, 1, 0)(2,1,1)[12] model to the original time series.






The time plot of the in-sample forecast errors shows that the variance of the forecast errors seems to be roughly constant over time. The histogram of the time series also shows that the forecast errors are roughly normally distributed and the mean seems to be close to zero. Therefore, it is plausible that the forecast errors are normally distributed with mean zero and constant variance. Since successive forecast errors do not seem to be correlated, and the forecast errors seem to be normally distributed with mean zero and constant variance, the seasonal ARIMA (2,1,0)(2,1,1)[12] does seem to provide an adequate predictive model for the monthly crude oil production.

## REFERENCES

[1] Alonso and Garcia (2012). "Time Series Analysis".Page 19.

[2] Akaike H,(1978) "Akaike's information criterion"

[3] Asemota, O.J. (2010). "Modelling Nigeria's crude oil Exports: State Space versus ARIMA model", Manuscript Submitted for Publication.

[4] Brockwell, P.J. and Davis, R.A. (1996) (Introduction to Time series and forecasting" Springer, New York. Section 3.3 and 8.3

[5] Dickey, D.A. and Fuller, W.A. (1981) "Likelihood Ratio Statistics for Autoregressive Process". Econometrics, 49(3), pp. 1057-1072

[6] Helmenstine, Anne Marie: URL: http//www.about.com/chemistry/petroleum definition, (August 4, 2014).

[7] Nigerian National Petroleum Commision (2016): Annual Statistical Bulletin

[8] Nigerian National Petroleum Corporation(2016): Industry History of nnpc.

[9] NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, August 4, 2014.

[10] Paul K (2012): Literature review of modern time series forecasting methods

[11] http://a-little-book-of-r-for-time-series.readthedocs.org/en/latest/src/timeseries.html, August 1, 2014

[12] R Package(Version 3.1.0)