PREDICTING MONTHLY RAINFALL OF UMUDIKE, ABIA STATE USING BOX AND JENKINS APPROACH

APJOCASR
www.akwapolyjournal.org
P-ISSN: 2536-6733 | E-ISSN: 2616-0773

Supported by:
tetfund
TERTIARY EDUCATION TRUST FUND
TETFund/DESS / POLY / IKOT-OSURUA/ ARJ /1-2)

[1]Udoudo, Unyime Patrick
Department of Statistics, Akwa Ibom State Polytechnic,
Ikot Osurua, Ikot Ekpene
udoudogeno@gmail.com

[2]Essien, Eduma Essien
Department of Statistics, Akwa Ibom State Polytechnic,
Ikot Osurua, Ikot Ekpene
edumaessien@gmail.com

Abstract

The Box and Jenkins method has been utilized to identify and fit a time series model to the monthly rainfall series of Umudike, Abia State, Nigeria. The data used in the study was obtained from the National Root Crop Research Institute (NRCRI) and covers the period between 1981 and 2020. The data analysis has revealed that the most suitable model for the series is SARIMA (0,0,0) x (1,1,0)12. A seasonal autoregressive component and a moving average component characterize this model. This model's two-year prediction indicates that the forecast is relatively stable. This means the predicted rainfall values will be similar to the historical values within the next two years. The study has also shown that using the SARIMA (0,0,0) x (1,1,0)12 model is reliable for modelling and forecasting rainfall in Umudike, Abia State, Nigeria. This information can be helpful for various stakeholders in the agricultural sector, enabling them to make informed decisions regarding crop production and management.

Keywords: *Rainfall, Time Series, Seasonal Arima, Forecasting.*

Introduction

It is crucial to understand the effects of climate change caused by national and anthropogenic factors and distinguish between them (Volker & Ingeborg, 2006). The global community is currently experiencing unfavourable climate conditions, including the gradual loss of rainforests in the tropics, the extinction of plant and animal species, changes in rainfall patterns, and global warming resulting from climate change. Climate change threatens social, political, and economic human development and survival (Olufemi et al., 20011). For many countries worldwide, including Nigeria, climate change is one of the biggest environmental threats to food production, water availability, forest biodiversity, and livelihoods (Nury et al., 2012). Developing countries in tropical regions are believed to be more severely affected than developed ones. In Nigeria, where the climate is tropical, there are significant climatic variations in different regions. The coastal areas experience high humidity and rarely exceed temperatures of 32°C (90°F), while the inland regions have two distinct seasons: a wet season from April to October with lower monthly temperatures and a dry season from November to March, with midday temperatures that rise above 38°C (100°F) but relatively cool nights, dropping as low as 12°C (54°F).

An essential aspect of improving agricultural productivity and human health is a better understanding of the potential climate change in Umudike. This can be achieved by carefully examining temperature data obtained from the meteorological unit of the National Root Crops Research Institute (NRCRI) in Umudike. An essential aspect of improving agricultural productivity and human health is a better understanding of the potential climate change in Umudike. This can be achieved by carefully examining temperature data obtained from the meteorological unit of the National Root Crops Research Institute (NRCRI) in Umudike. Time series analysis of weather data, such as rainfall, can be a valuable tool for investigating its variability pattern and predicting short- and long-term changes. Time series analysis and forecasting are significant tools in numerous meteorological applications to study trends and variations in variables like rainfall, humidity, temperature, stream flow, and other environmental parameters.

Several studies have used time series analysis to detect rainfall and runoff pattern changes. Some of these studies include Langu (1993), as cited by Nail and Momani (2009), who used time series analysis to detect changes in rainfall. Etuk et al. (2013) modelled the monthly rainfall as measured in Port Harcourt, Nigeria, as a seasonal ARIMA model, while Osarumwense (2013) modelled the quarterly rainfall data as a seasonal ARIMA model. Olofintoye and Sule (2010) fitted a trend line indicating a positive rainfall trend. Haboya and Igbinedion (2019) researched the capacity of linear and non-linear regression techniques. They found that the artificial neural network (ANN) had a higher coefficient of determination R2 than the multiple linear regression (MLR). Gnanasankaran and Ramajoaj (2020) conducted various linear regression models to predict rainfall using Nigerian meteorological data. Refonna, Lakshini, Raza Abbas, and Mohammed Razinilla (2020) predicted rainfall using machine-learning techniques and linear models.

This study aims to investigate whether there is a trend in the rainfall pattern of Umudike from 1981 to 2020, model the trend (if any) identified using time series analysis, and use the identified model to predict future rainfall occurrence in Umudike. The study's importance lies in its theoretical contribution, which fills gaps in the literature, and its practical implications, which will guide farmers in Umudike to predict the amount of rainfall and make informed decisions about what farming methods to adopt.

Methodology

The data for this study is secondary data and was collected from the meteorological department of the National Root Crop Research Institute (NRCRI), Umudike. A significant limitation in data collection is data confidentiality. This effect presented some problems in gathering enough facts regarding this study.

The time series method of using the Box and Jenkins in methodology was adopted, and its accompanying assumptions, like meeting the stationarity condition, were adequately justified.
There are four-time series components: a skeletal trend, seasonal variation, cyclical variation, and irregular variation.

The Box-Jenkins method

The Box-Jenkins method is a methodology which uses a variable's past behaviour to select the forecasting model from a general class of models. Three stages are involved in this methodology; these include identifying the tentative model, determining the model'sparameters, and applying the model.

They are identifying the tentative model involved in making sure that the variables are stationary and using plots of the dependent time series's autocorrelation and partial autocorrelation functions to decide which (if any) auto-regressive or moving average component should be used in the model.

Model Identification

The identification stage involves determining tentative values of *p,d, and q* and the *P, D, and Q* sets using the linear least squares method. In the Identification stage, a stationary or a weakly stationary situation is obtained by differencing and transforming the data if needed. Then, the ACF and PACF plots suggest possible models by determining the orders *p* and *q* in the seasonal ARIMA *(p,d,q)(P, D, Q)* model. The goodness of best models could be evaluated using the mean square error (Residuals) MSE or the Akaike Information Criterion.

- **Autocorrelation Function (ACF)**

It is the similarity between observation as a function of the time separation between X_t and X_{t+k}at lag k. It is defined by;

$$\rho_k = \frac{E[(X_t - \mu)(X_{t+k} - \mu)]}{\sqrt{E[(X_t - \mu)^2 E[(X_{t+k} - \mu)^2]}}$$
$$= \frac{Cov\ (X_t, X_{t+k})}{\sqrt{Var\ (X_t) Var\ (X_{t+k})}}$$
$$= \frac{Cov\ (X_t, X_{t+k})}{\sigma_x^2}$$

Therefore, $\rho_k = \frac{r_k}{r_0} = \frac{R_k}{R_0}$

Where $\sigma_x^2 = r_0 = R_0$for stationary process, thus the autocorrelation at lag K is

$$\rho_k = \frac{r_k}{r_0}$$

Where $\rho_0 = 1, 2, \dots$ k = 0 $\pm$ 1 $\pm$ 2 $\pm$ 3 $\pm$

Partial Autocorrelation Function (PACF)

Given a process $X_t, t \in Z$, the partial autocorrelation function is said to be the correlation between X_t and X_{t+k}. After removing their mutual dependency on the intermediate or intervening variables X_{t+1}, $X_{t+2}, \dots,$ X_{t+k-1}. This is denoted as $\emptyset_{kk}$ and it is defined as:

$$\phi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1} \phi_{k-1} \rho_{k-j}}{1 - \sum_{j=2}^{k-1} \phi_{k-1}, \rho_{k-j}} k = 2,3$$

and $\emptyset_k . j = \emptyset_{k-1}, j - \emptyset_{kk} \emptyset_{k-1,k-1}$ $\qquad j = 1,2,3, \dots, k-1$

- **Autoregressive Integrated Moving Average Model (Arima)**

In practice, most time series are non-stationary. In order to fit a stationary model, the method of differencing is applied. This method is particularly useful for removing trend in the series. For non-seasonal data, first order differencing is usually sufficient to attain apparent stationarity. Here a new series say $(Y_1, Y_2, \dots, Y_N)$ is formed from the original observed series say $(X_1, X_2, \dots, X_N)$ by;

$$Y_t = X_t - X_{t-1} = \nabla x_t \text{ for } t = 1,2,3, \dots N$$

Occasionally, second order differencing is required.

Giving the ARIMA model

$$X_t = \emptyset_1 X_{t-1} + \cdots + \emptyset_P X_{t-P} - \theta_1 e_{t-1} - \cdots - \theta_q e_{t-q} + e_t$$

By backshift, the equation becomes

$$\emptyset(B)X_t = \theta(B)e_t$$

If X_t is replaced by $\nabla^d X_t$then we have a model capable of describing certain types of non-stationary series, such a model is called an integrated model because the stationary model that is fitted to the differenced data has to be summed or integrated to provide a model for the original non-stationary data written as

$$W_t = \nabla^d X_t = (1 - B)X_t$$

The dth difference of X$_t$ is said to be an ARIMA process of order (p,d,q).

Choice of Model and Order Through ACF and PACF

- **AR(1):** ACF decays exponentially and PACF cuts-off or spikes at lag 1. The model is given as

$$X_t = \emptyset_i X_{t-1} + e_t$$

- **AR(2):** ACF has a sine-wave pattern or a set of exponentially decays, PACF cuts-off or spikes at lag 1 and lag 2 and no correlation for other lag.

$$X_t = \sum_{t=1}^{P} \emptyset_i X_{t-i} + e_t$$

$$\text{i.e. } X_t = \emptyset_1 X_{t-1} + \emptyset_2 X_{t-2} + e_t$$

- **MA(1):** ACF cuts-off at lag 1 and no correlation for other lags and PACF decays exponentially, the model is

$$X_t = e_t - \theta_i e_{t-1}$$

- **MA(2):** ACF cuts-off at lag 1 and 2 and no correlation for other lags and PACF exhibits a set of sine-wave pattern or exponentially decays.

$$X_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2}$$

This involves finding the appropriate values of (p,d,q) for the ARIMA process for the series. This can be achieved by the use of Auto-correlation Function (ACF) Correlogram and Partial Auto-correlation Function (PACF) Correlogram. The table below summarizes how sample ACF could be used for model identification.

Table 3.1: Summary of Sample ACF for Model Identification

S/No	Shape	Indicated Model
1	xponential decaying to ero	AR model, use the PACF plot to identify the order of the AR model.
2	lternating positive and egative decaying to zero	AR model, use the PACF plot to identify the order
3	ne or more spikes, rest re essentially zero	MA model, order identified by where plot becomes zero
4	ecay, starting after a ew lags	Mixed AR and MA model
5	ll zero or close to zero	Data is essentially random
6	igh values at fixed	Include seasonal AR term
7	o decay to zero	Series is not stationary

Estimation and Diagnostic Checking

Having identified p and q values, the autoregressive and moving average parameters were estimated using R software. After choosing a particular ARIMA model, the next thing is to check whether the model fits the data appropriately since there is a possibility of choosing another ARIMA model that might do the same work.

A simple test of the adequacy of the chosen model is to see if the residual estimated from the model is white noise.

If the residuals are not white noise, then there is a need to start again in the identification and specification of another model. The iteration continues until the suitable model is fitted to the data.

If the fitted model is adequate, the residuals should be approximately white noise.

So, we should check if the residuals have zero mean and are uncorrelated. The key instruments are the time plot and the ACF and PACF of the residuals. The theoretical CF and PACF of white noise processes take value zero for lags, so if the model is appropriate, most of the sample ACF and PACF coefficients should be close to zero. Out 95% of these coefficients should fall within the non-significant bounds. e degrees of freedom of the statistic consider the number of estimated parameters; thus, the statistic test under H0 follows a distribution approximately.

If the model is appropriate, we expect the residuals' correlograms (superficial and partial) to depart from white noise, suggesting the reformation of the model. e statistic used was proposed by Box and Pierce (1970), known as Box-Pierce, and is used to test the model's accuracy using autocorrelation.

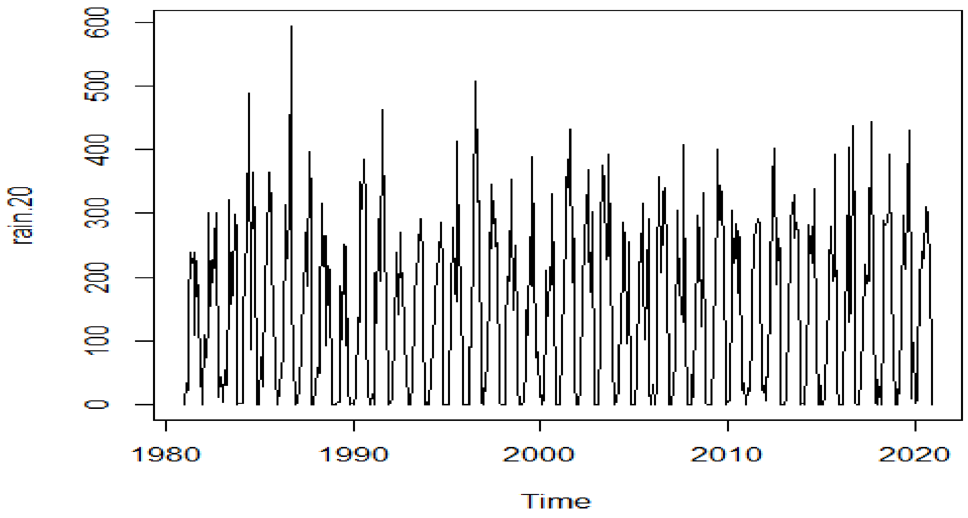
The form of the statistic is given by:

$$Q = n\sum_{j=1}^{T} Pj^2$$

The critical region is given by $\sim \chi^2_{a.T-p-q}$ where T is the default number of lags in the ACF of the residual p=AR parameter, q=MA parameter $\hat{P}_j$=Correlogram.

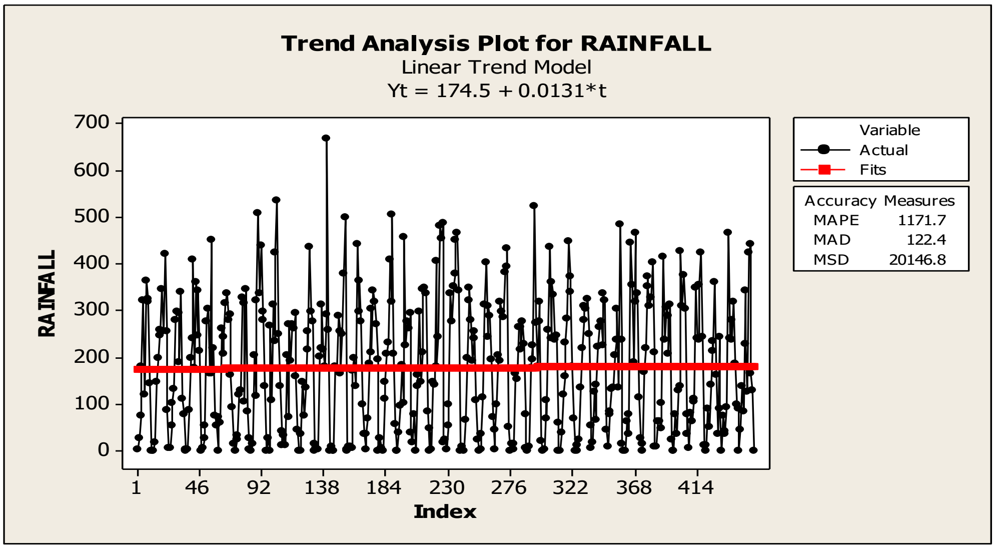
3. Results and Discussions

The plot below shows the time plot as well as the ACF and PACF of the time series data on monthly rainfall in Umudike as recorded by the National Root Crop Research Institute, Umudike.

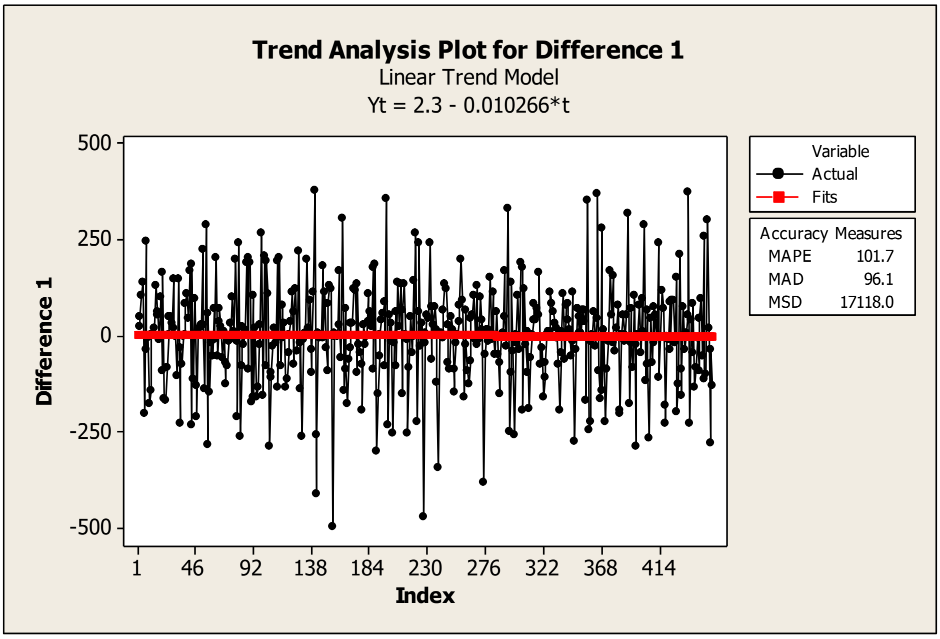


3.1 Model Identification

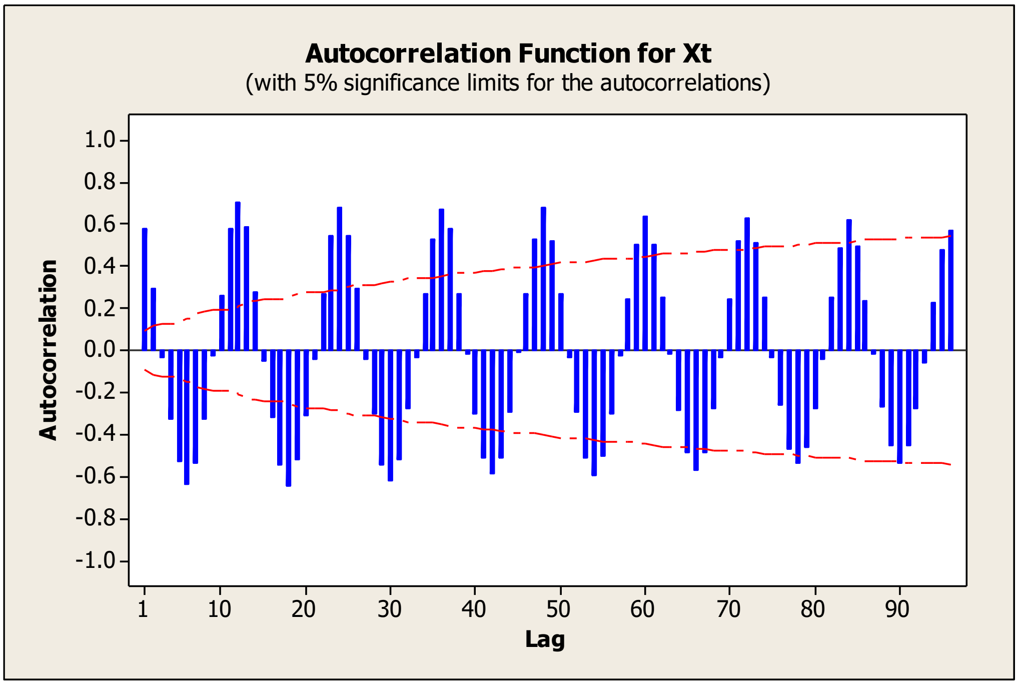
The trend analysis of the time plot for the original data suggests a possible presence of a positive small (almost negligible) trend in the time series data. In order to remove the trend, the data was differenced at lag 1 and 2 and then the time plot and trend analysis was done.

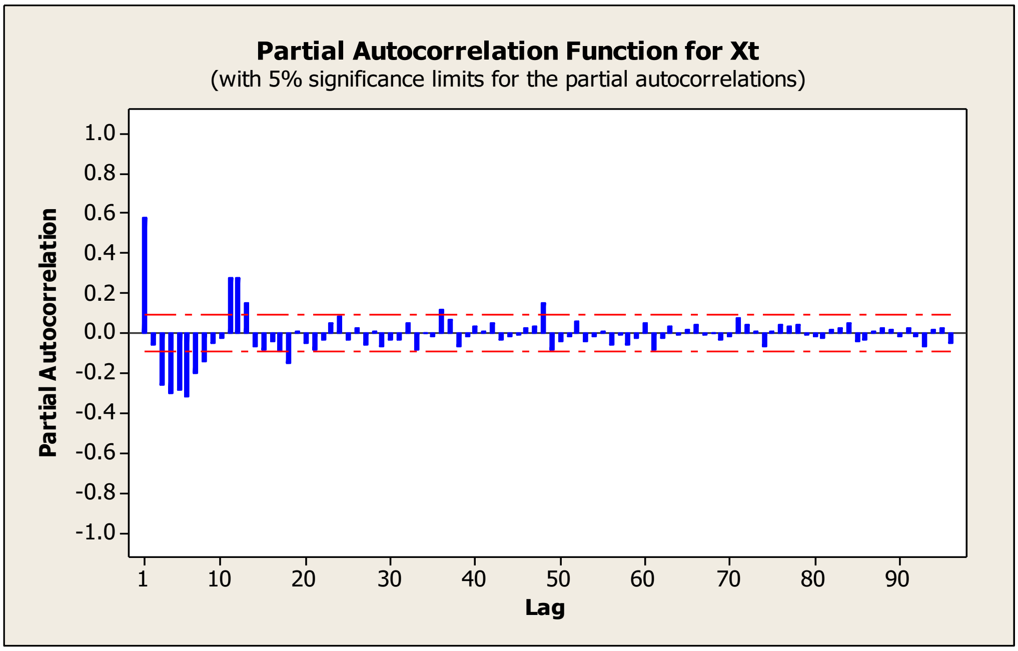


In examining the trend analysis of difference one and two, the differencing at lag one had the minimum mean square error and as such was selected. It is assumed that the secular variation has also been isolated by differencing at lag 1 and 2, thus there may be no more presence of trend in the time series data. The trend analysis of difference one is shown below.



The ACF of the original as shown below has a sin qua sin movement with a first significant cut-off at lag 1 and 2. This suggests the presence of seasonality in the model. The PACF also as shown had a significant cut-off at Lag 1 and after every 12 lags. This also suggests the presence of a seasonal variation in the model.

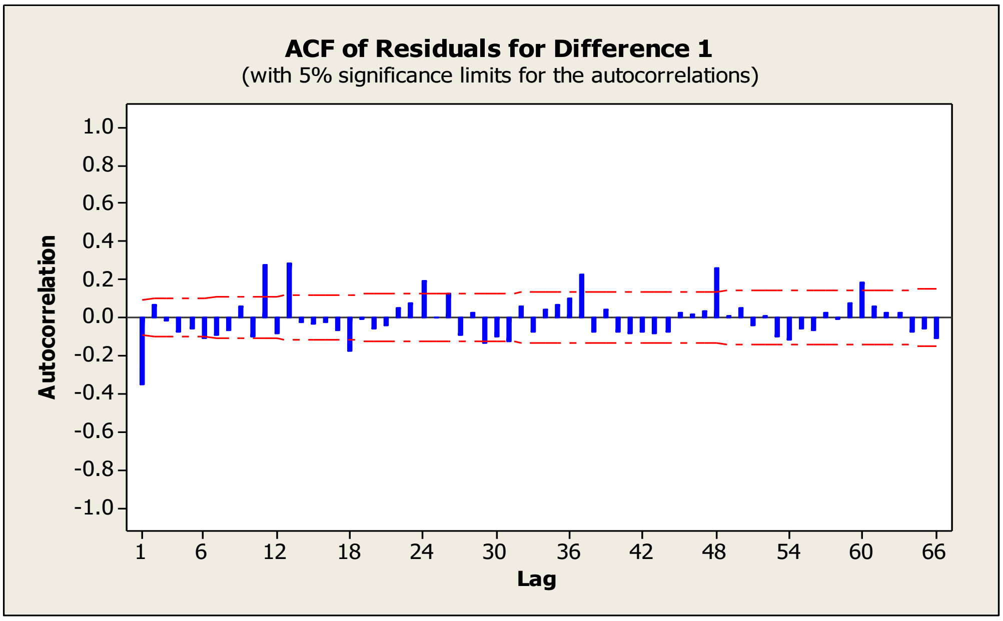


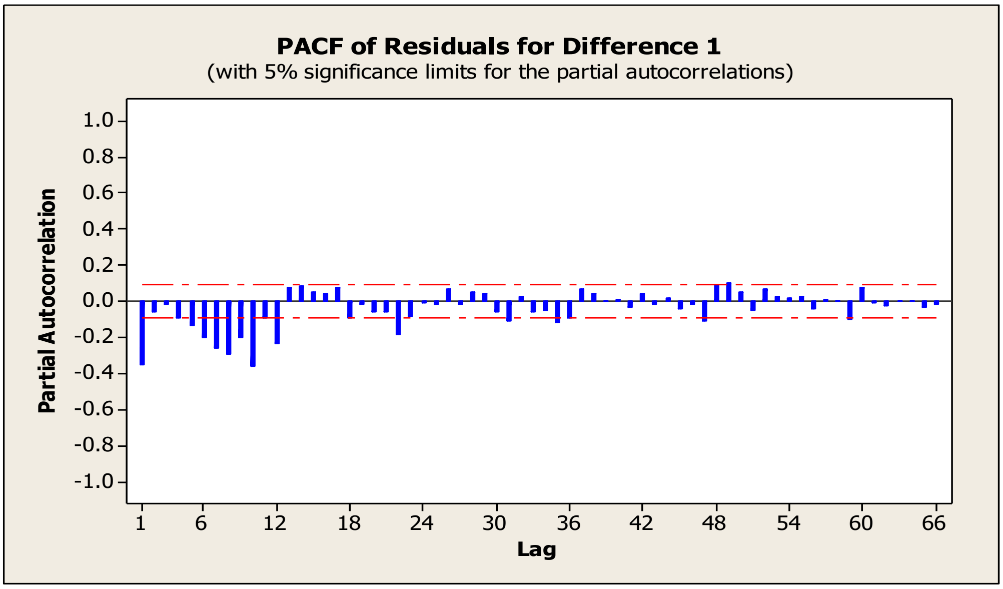


A SAR (1) estimate 0.9609, which is closer to one strongly suggests that the model is non-stationary at this stage and might need to be differenced at the seasons. The ACF and PACF of the differenced SAR(1) preliminary analysis of the monthly rainfall up to the lag 90. Observe that ACF (left panel) seems to cut off after the first season suggesting a model with Q = 1. The PACF also have a significant cut-off at lag one suggesting an initial model gaze with P = 1.

A SAR (1) estimate 0.9609, which is closer to one strongly suggests that the model is non-stationary at this stage and might need to be differenced at the seasons. The ACF and PACF of the differenced SAR(1) preliminary analysis of the monthly rainfall up to the lag 90. Observe that ACF (left panel) seems to cut off after the first season suggesting a model with Q = 1. The PACF also have a significant cut-off at lag one suggesting an initial model gaze with P = 1.

So far, a detected seasonal effect s, an indicative Q = 1, P = 1from the ACF and PACF respectively, and a parsimonious D = 1 from the first seasonal difference. Thus, tentative models obtainable from the ongoing preliminary analysis are shown in the table below.





MODEL	MSE
SARIMA(0,0,0)(1,1,1)	0.5919
SARIMA(0,0,0)(2,1,1)	0.5931
SARIMA(0,0,0)(3,1,1)	0.5902
SARIMA(0,0,0)(1,1,0)	0.5897
SARIMA(0,0,0)(5,1,1)	0.6667
SARIMA(0,0,2)(2,1,1)	0.5957
SARIMA(0,1,2)(1,1,1)	0.6130
SARIMA(1,1,2)(1,1,1)	0.6064
SAMIRA(2,1,2)(1,1,1)	0.6064

Table 1: SARIMA of NRCRI Monthly Rainfall (1981-2020)

Based on the MSE of the various indicative Seasonal ARIMA models displayed in the table above, the SARIMA (0,0,0) $(1,1,0)_{12}$ could be chosen as the best fitted model for the NRCRI monthly Rainfall.

3.2 Estimation

Using MINITAB software, the following model parameters were estimated.

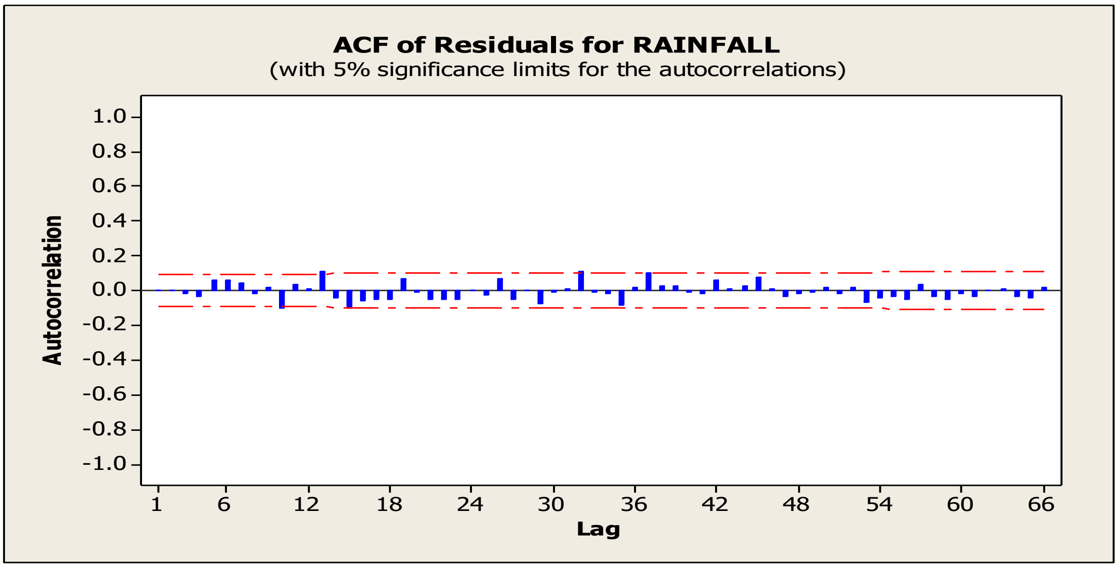
SAR = $\emptyset = 0.0215$

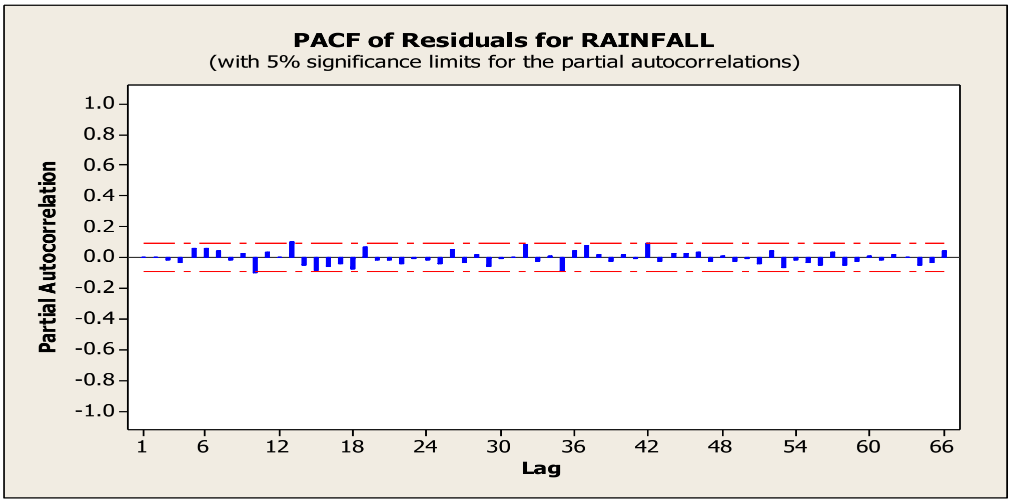
SMA = $\theta = -0.4283$

Having identified a tentative model with the above estimated parameters, the model was diagnosed to ascertain its adequacy.

3.3 Diagnostics Checking

To check the adequacy of this model, the ACF and PACF of the residual was plotted to check for randomness of the residual for the SARIMA (0,0,0)(1,1,0). It can be observed that all the values of ACF and PACF were seen within the 95% confidence bound of the plots.

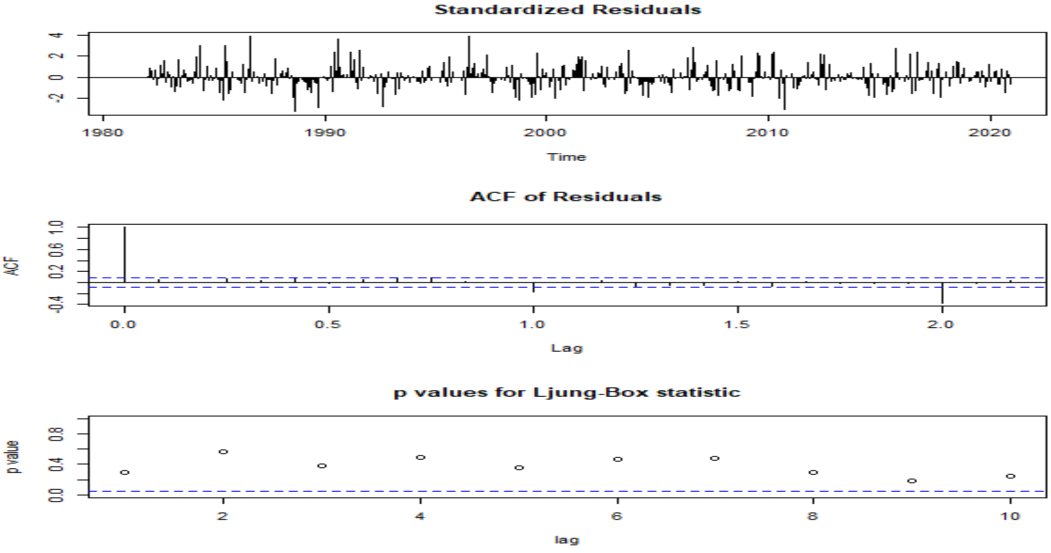




The plot of the ACF and PACF of the residual below suggests that no more pattern was left. It also indicates evidence of randomness of the auction goods thus implying that SARIMA (0,0,0)(1, 1, 0) is a good model for the data. Usually, we compute the Box-Pierce estimate with the hypothesis Ho: the model is adequately fit the data and H_1: the model is not adequate for the fitted data. The modified Box-Pierce (Ljung-Box) Chi-Square statistic

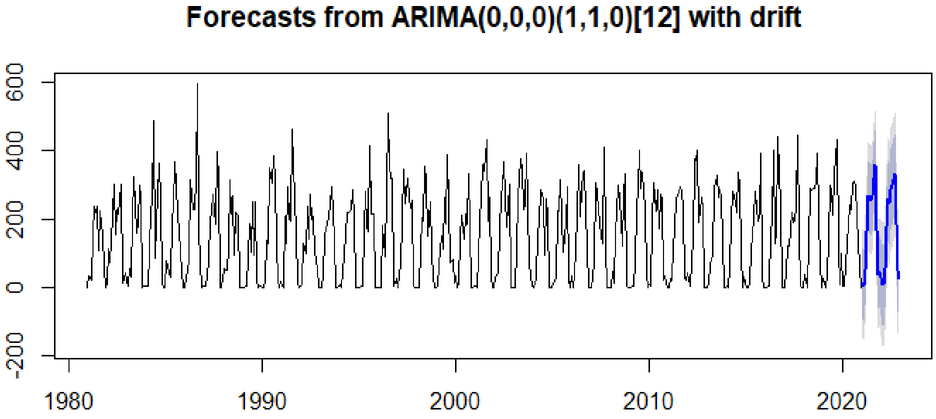
Lag	12	24	36	48
Chi-Square	10.8	28.0	43.5	56.5
DF	6	18	30	42
P-Value	0.094	0.062	0.053	0.067

Since all p-value for sample size 12, 24, 36 and 48 is greater than 0.05 as seen above, we accept Ho and conclude that the model is appropriate.



3.4 Forecasting

Using the identified model parameter in forecasting one or more future time steps, the following results were obtained and giving in the plot below. Having confirmed that the model is appropriate for forecast, a two year (2021-2022) forecast was done with the grey lines representing the two years confidence limits at 95% Limit.



4.1 Summary

The study examines 1981-2020 monthly rainfall data for Umudike collected using a Metrological instrument in the NRCRI Umudike at latitude 050, 29’N and longitude 070, 33’E(122M above sea level). The time plot of the data shows the possible presence of a slight trend and the presence of seasonality. The classical Box and Jenkins Time series methodology was employed with its indicative ACF and PACF identification guide. The SARIMA(0,0,0)(1,1,0)12 model adequately forecasted the monthly rainfall from 1981 to 2020. Verification of the model was performed using the 1981-2020 time period. In the end, it was recommended that carefully studying mathematical models could help track future rises in monthly rainfall.

4.2 Conclusion

The Box-Jenkins ARIMA methodology is a valuable technique that can help decision-makers establish better strategies and set priorities for equipping themselves against upcoming weather changes. The rainfall data fitted by this procedure for the NRCRI station shows that it is possible to predict the evolution of rainfall in Eastern Nigeria based on data collected over the past 39 years.

Based on the best-suited model, the monthly rainfall for the next two years is more stable than the reference period 1981-2020. Of course, this approach has shown dependable results, and a general recommendation is that careful understudying of mathematical models like the ARIMA could help track future rises in monthly rainfall for relatively short time intervals. As we advance, the uncertainty about the predictions grows so that the result might become indecisive.

References

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

Etuk, E. H., Mofatt, I. U & Chims, B. E. (2013), Modelling Monthly Rainfall Data of Port Harcourt, Nigeria by Seasonal Box-Jenkins Methods, *International Journal of Sciences,* 2 (7), 60-67.

Gnanasankaran, N. & Ramara, E. (2020). Multiple linear regression model to predict rainfall using Indian Meteorological data. *International journal of advanced science and technology.* Vol.29, No. 08s, pp 746-758.

Haboya, R., & Igbinedion, O. (2019) Performance of multiple linear regression (MLR) and Artificial neural network (ANN) for the prediction of monthly maximum rainfall in Benin City, Nigeria. *International journal of engineering science and application,* Vol. 3, No. 1.

John C., George, U. & Chukwuemeka, O. S. (2013), Time Series Analysis and Forecasting of Monthly Maximum Temperatures of South Eastern Nigeria, *IJIRD,* 3(1), 167-171.

Nail, P. E. & Momani, M (2009): Time Series Analysis Model for Rainfall Data in Jordan; A case Study for Using Time Series Analysis, *American Journal of Environmental Science* 5(5): 599-604. Science Publication.

Olofintoye, O. O. & Sule, B. F. (2010): Impact of Global Warming on the Rainfall and Temperature in the Niger Delta of Nigeria. *USEP-Journal of Research Information in Civil Engineering,* 7.2, pp 33-48

Oluwafemi, S.O., Femi, J.A. & Oluwatosi, T.D. (2010), Time Series Analysis of Rainfall and Temperature in South West Nigeria, *The Pacific Journal of Science and Technology,* 11(2), 552-564.

Osarumwense, O. I (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, 4,1 pp. 1-4.

Refonna, J., Lakshmi., Abbas, R. & Kazuillha, M. (2019). Rainfall prediction using Regression Model. *International Journal of Recent Technology and Engineering (IJRTE)* ISSN:2277-3878, Vol.8, Issue-2s3.

Volker, W. & Ingeborg, H. (2006), The Climate Station of the University of Hohenheim: Analyses of Air Temperature and Precipitation Time Series since 1878; *International Journal of Climatology,* 26, 113-138.