ARIMA MODEL FOR FORECASTING OF MONTHLY RAINFALL AND TEMPERATURE IN THE LAKE VICTORIA BASIN

Original Article

Abstract

Global warming has given rise to erratic and disruptive weather pattern in several regions of the world including the area surrounding the Lake Victoria basin. Likewise, economic activities associated with Lake Victoria and its Basin such as agriculture, fishing, mining and transportation are significantly affected by this climatic changes. The primary cause of negative impact that stems from this changes is lack of reliable information that can be used to predict and address the climatic variations within the basin. The objective of this research is to identify a suitable time series model that can be used in analysing and predicting this weather variations and pattern around the Lake Victoria basin. This research uses Box jenskin methodology to build ARIMA(2,0,1) model for rainfall pattern around Lake Victoria basin. The data is obtained from three Kenya Meteorological Department weather stations as secondary data from the years 2008 to 2014. In this research, data from the years 2008 to 2010 was used to estimate the values for the years 2011 to 2013. The relationship from the research showed a strong positive relationship which indicates high level of accuracy on predictability by the model.

Keywords: ARIMA, Box-jenskin,

xxxxx

1 Introduction

Climate change is described as changes in the state of climate that can be identified by statistical variables and tend to persist for a long time [4]. Global climatic changes, known as global warming, is caused by a high concentration of carbon and fluorine related gases in the atmosphere(GHG). The gases have a profound effect on the environment and affects climatic distribution patterns and hence regional climate changes[3]. It is projected that climate change will increase and have an adverse effect on humanity and nature [2]. The environment that supports Lake Victorian basin ecosystem is becoming fragile from weather changes which are extreme and erratic. Such fluctuations have negative impact on socio-economic activities that takes place along the Lake Victoria Basin.

Absence of precise, reliable and consistent information from weather forecasters for these temperature and rainfall distribution pattern creates uncertainty and lack of anticipation by policy makers, policy implementers and the general inhabitants of Lake Victoria Basin who directly depend on its environment. Examples of extreme weather conditions that was not predicted includes the expected El Nino rainfall and temperature occurrences in Kenya from December 2014 to February 2015, predicted rainfall in December 2015 that never materialized and the devastating drought from the year 2009 to the year 2010 that was never anticipated. This project will use rainfall data from the years 2007 to 2014 from selected Kenya Meteorological Department's stations to identify a suitable Time series model that can give precise weather forecast for the Lake Victorian Basin. The most commonly used Time series model is the ARIMA model developed by Box and Jenkins [1]. ARIMA models, also called the Box-Jenskin models, are models that relate the present value of a series to the past values and past prediction errors.

ARIMA models includes Auto-regressive terms,moving average terms and differencing operation. ARIMA stands for a series which needs differencing to be made stationary. Lags of stationary series are called Auto-Regressive(AR) terms while lags of forecast errors are called Moving Average(MA) terms.

1.1 Basic Concepts

1. *A time series is said to be strictly stationary if the joint probability distribution of the process does not change when shifted in time i.e., If we take X_t to be a stochastic process and*

$$X_{t_1}, X_{t_2}, X_{t_3}, \cdots, X_{t_n} \tag{1.1}$$

is the same as

$$X_{t_{h+1}}, X_{t_{h+2}}, X_{t_{h+3}}, X_{t_{n+}} h \,\forall\, t_i \in R. \tag{1.2}$$

In stationary series, the mean and variance do not change with time, It has no periodic variations, has no trend and its autocorrelation is constant. A time series is called 2^{nd} order stationary or weakly stationary if it has a constant mean and its auto covariance function is independent of time but dependent only on the distance between the variables and its mean is finite,i.e if

$$E(X) = \mu < \infty \,\forall\, t \in R$$

Where $E(X)$ is the expectation of the X
Taking X at times t as X_t it will follow that

$$E(X_t) = \mu_t \, and \, E(Xt + h) = \mu_{t+h}$$

applying the covariance equation for two variables XY given by

$$cov(XY) = E(X - \mu_x)(Y - \mu_y) = \sigma$$

for the variables X_t and X_{t+h} we obtain

$$E(X_t - \mu_t)E(X_{t+h} - \mu_{t+h}) = \sigma(h) \tag{1.3}$$

2. ***Auto correlation function*** $\rho(h)$. *Auto correlation function is a measure of how much(significant) the present variables are correlated with the past variables at a given lag h and helps in determining how far back the variables are correlated. The values of auto correlation varies between +1 and -1. If the covariance of* X_t *and* X_{t+h} *is given by* $\sigma(h)$ *and their respective variances as* $V(X_t) and V(X_{t+h})$ *respectively Then*

$$\rho(h) = \frac{\sigma(h)}{\sqrt{V(X_t)V(X_{t+h})}} \tag{1.4}$$

$\rho(h)$ *represents the auto correlation function (ACF) of a time series at a time lag of h between the variables* X_t *and* X_{t+h} *and this varies between -1 and +1*

3. ***Partial autocorrelation function*** *(PACF) are a measure of correlation between variables* X_t *and* X_{t+h}*. if there is a large set of lags in between that is making the auto correlation difficult to establish, Partial auto correlation function gives the partial correlation within its lagged values thereby handling shorter lag values. The PACF is used in data analysis to identify the extent of a lag in ACF. If* Φ_{hh} *represent the coefficient of partial regression of the* r^{th} *order auto regression, Then*

$$X_{t+h} = \Phi_{h_1}X_{t+h-1} + \Phi_{h_2}X_{t+h-2} + \cdots + \Phi_{hh}X_t + e_{t+h}$$

where e_{t+h} *is a normal error term. Multiplying* X_{t+h} *and* X_{t+h-j} *and finding the expectation we obtain its covariance at lag h given by*

$$\sigma(h) = E(X_{t+h}, X_{t+h-j})$$

The covariance at lag h is then divided by covariance at lag 0 to find the partial auto correlation function

$$\rho(h) = \Phi_{h_1}\rho(j-1) + \Phi_{h+2}\rho(j-2) + \cdots + \Phi_{hh}\rho(j-h)$$

4. ***Moving average process****(MA). Suppose* e_t *is a white noise(serially uncorrelated random variables with zero mean and finite variance), then the process*

$$X_t = \theta_0 e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \theta_3 e_{t-3} + \cdots + \theta_q e_{t-q}$$

is the moving average process of order q and can be represented as

$$X_t = \sum_{j=0}^{q} \theta_j e_{t-j} \tag{1.5}$$

$\theta_1, \theta_2, \theta_3, \cdots, \theta_q$ *are the parameters of the moving average process with q being the maximum order.*
The mean μ *of MA process given by* $E(X_t)$ *is zero since* $E(e_t) = 0$ *The variance of MA process is given by*

$$var(X_t) = Var(\sum X_t)$$

Taking the variance of e_t *as* σ *which is a constant then*
$var(X_t) = \sigma^2 \sum \theta_j \theta_j$ *at lag* $h = 0$.

$$var(X_t) = \sigma^2 \sum \theta_j^2 \tag{1.6}$$

The covariance of X_t*, with*

$$E(X_t) = 0$$

will be

$$E(X_t; X_{t+h}) = \sum \theta_j \theta_{j+h} E(e_t e_{t+h})$$

hence the covariance is

$$\sigma(h) = \sigma^2 \sum \theta_j \theta_{j+h} \tag{1.7}$$

The auto correlation coefficient for an MA process is give by

$$\rho(h) = \frac{\sigma^2 \sum \theta_j \theta_{j+h}}{\sum \theta_j^2} \tag{1.8}$$

5. ***Auto regressive process***. *A time series described by the process*

$$X_t = \delta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} + \cdots + \phi_p X_{t-p} + e_t \tag{1.9}$$

Where $\phi_1, \phi_2, \phi_3, \phi_p$ *are some constants representing the parameter of the series, p represents the order of the series and* e_t *represents normally, identically and independently distributed random error term with mean* $\mu = 0$ *and variance* σ^2*.*
An autoregressive process that is stationary has the absolute values to the solution of the equation $\phi(B) = 0$ *and lie outside the unit circle in the complex plane where B is a backward shift operator such that*

$$B(X_t) = X_{t-1}$$

$$B^2 X_t = B(BX_t) = X_{t-2}. \tag{1.10}$$

The AR Model and the AR polynomial.

$$\Phi(B) = 1 - \phi_1 B + \cdots + \phi_p B^p \tag{1.11}$$

For AR(1) we have

$$X_t = \delta_1 X_{t-1} + \omega_t \tag{1.12}$$

Hence

$$(1 - \phi_1 B) X_t = \delta + \omega_t$$

Denoted as

$$\Phi(B) X_t = \delta + \omega_t$$

Where $\omega_t \sim N(0, \sigma\sigma^2)$ *and* $\Phi(B) = 1 - \phi_1 B$ *is an AR polynomial.*

6. ***Auto regressive moving average process***. *The combination of autoregressive process (AR(p)) and moving average process (MA(q)) from a stochastic model in develops to get Auto Regressive Moving Average (ARMA) model. The model represents a stationary time series process. The ARMA model is represented as*

$$\phi_1 + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} = \theta_1 e_1 + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} \tag{1.13}$$

Using a backward shift operator

$$\phi(B) X_t = \theta(B) X_t \tag{1.14}$$

where $\phi(B) = 1 - \phi_1 B + \phi_2 B^2 + \cdots + \phi_p B^p$ *and* $\theta(B) = 1 - \theta_1(B) + \theta_2 B^2 + \cdots + \theta_q B^q$
These equations are polynomials of degree p and q that forms an ARMA model with parameters p, q*. For stationariness, both the absolute values of the solutions to the polynomials must lie outside the unit circle.*

7. ***Differencing*** *A non stationary time series can be made stationary by either linear filtering or differencing method. In this research, the differencing method is applied such that*

$$\nabla X_t = X_t - X_{t-1} \tag{1.15}$$

which representing a first order differencing and

$$\nabla(\nabla X) = \nabla^2 = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) \tag{1.16}$$

representing a second order differencing

8. ***Auto regressive integrated moving average process****.*
If a time series has an ARMA model with a trend that is not stationary, the model is integrated by differencing to make it stationary.
Let ∇^d represent d times differencing to produce a stationary model, by using a backward shift operator function

$$\nabla = (1 - B) \tag{1.17}$$

this implies

$$\phi(B)\nabla X_t = \theta(B)e_t \tag{1.18}$$

can be represented as

$$\phi(B)(1 - B)^d X_t = \theta(B)e_t \tag{1.19}$$

This is Auto Regressive Integrated Moving Average (ARIMA) model of order p, d, q.

9. ***Ljung-Box test;***
A test statistics on the residuals of ARIMA model used in confirming weather the data used in the research are random in nature or not. Generally, the residuals tests the hypothesis.

H_o : *The data is random*
H_1 : *The data is not random*

The test statistics is given by

$$Q_{LB} = n(n+2)\sum_{j=1}^{h} \frac{P^2}{n-j} \tag{1.20}$$

Where n is the sample size, p is the auto correlation at lag j and h is the number of lags. The hypothesis on the test of randomness is rejected when

$$Q_{LB} = X^2 \tag{1.21}$$

where X^2 is a chi-square distribution.

2 ARIMA MODEL

The research developed ARIMA model using secondary data that was collected from Kenya meteorological department.
Analysis was done by using Box jenskin methods of

1. *model identification*
2. *parameter estimation*
3. *data validation.*

2.1 Identification of the model

At this stage data is smoothed by using the moving average(MA) process and the central moving average(CMA) process. Smoothing is done so that a clear signal of patterns and trends in the time series is produced by removing irregular roughness. A twelve period moving average is used to center the twelve months of the year and six months central moving average is used as the average of the twelve months. The central moving average helps us to see the trend from the data.
The table below represents monthly rainfall with their moving average(MA) and central moving average(CMA) from the years 2008 to 2010.

Time	year	month	Rain	year	month	MA(12)	CMA(6)
1	2008	1	27.178	2008	1		
2	2008	2	68.072		2		
3	2008	3	146.05		3		
4	2008	4	164.338		4		
5	2008	5	150.368		5		
6	2008	6	101.09		6		
7	2008	7	83.312		7		
8	2008	8	86.36		8		
9	2008	9	81.534		9		
10	2008	10	260.069		10		
11	2008	11	144.526		11	154.9823	153.8938
12	2008	12	546.862		12	161.671	151.9162
13	2009	1	107.442	2009	1	159.2157	148.9166
14	2009	2	38.608		2	149.86	148.5446
15	2009	3	33.782		3	155.0882	146.9813
16	2009	4	227.076		4	151.003	144.2115
17	2009	5	101.346		5	145.4362	137.7557
18	2009	6	34.29		6	141.1393	131.8441
19	2009	7	31.75		7	140.6737	17.0786
20	2009	8	80.772		8	156.6122	124.1244
1	2009	9	272.796		9	138.9168	119.8003
22	2009	10	47.752		10	135.6995	117.6867
23	2009	11	105.918		11	105.8122	115.824
24	2009	12	188.214		12	104.0553	117.9437
25	2010	1	86.36	2010	1	107.7807	120.2569
26	2010	2	83.31	2010	2	119.9938	120.4474
27	2010	3	180.34	2010	3	126.3438	119.6673
28	2010	4	303.276	2010	4	124.1213	118.0616
29	2010	5	74.676	2010	5	122.6608	115.8754
30	2010	6	16.764	2010	6	120.65	114.1895
31	2010	7	7.62	2010	7	120.2478	113.1127
32	2010	8	75.946	2010	8	109.1142	111.6857
33	2010	9	139.192	2010	9	114.5328	112.3286
34	2010	10	112.776	2010	10	115.1043	111.5939
35	2010	11	112.776	2010	11	108.8178	109.8386
36	2010	12	112.776	2010	12	110.8595	110.8595

Rainfall moving average and central moving average

2.1.1 Time series plot of rainfall data

A plot of time series rainfall data indicates seasonal behaviour from the moving average as time increase weather the trends are increasing, decreasing or are constant. This plot can help us choose between multiplicative and additive methods and also show us the general trend of the series which in this case is a non increasing trend.

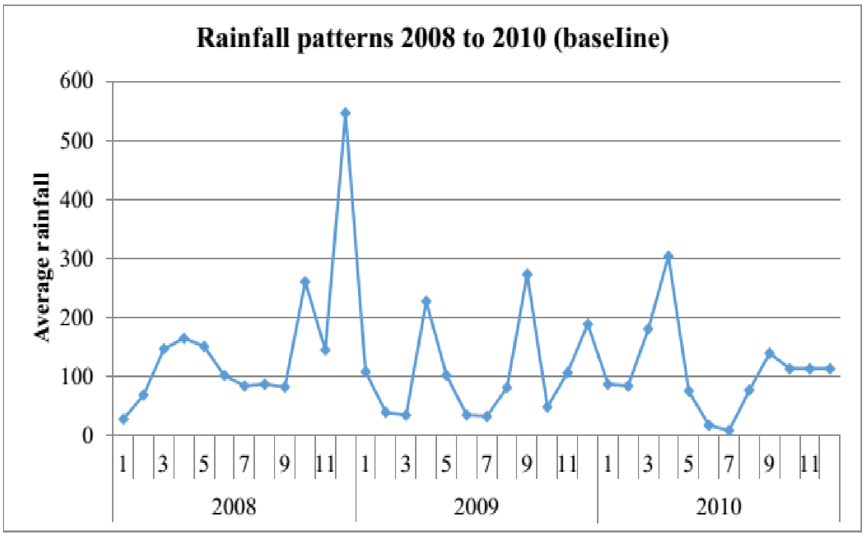

Figure 1: Time series plot for rainfall data

2.1.2 Parameter estimation

Auto Correlation Function plot *The two figures below illustrate autocorrelation function and Partial Autocorrelation functions respectively. As both ACF and PACF show significant values, we* *assume that an ARIMA model will serve our needs. The ACF can be used to estimate the MA-part, i.e. q-value, the PACF can be used to estimate the AR-part, i.e. p-value. To estimate a model-order we look at;*

1. *Whether the ACF values die out sufficiently*
2. *Whether the ACF and PACF show any significant and easily interpretable peaks at certain lags.*

ACF and PACF might suggest not only one model but many from which we need to choose from after considering other diagnostic methods. The most obvious model seems to be ARIMA (4,0,2) as ACF values die out at lag 4 and PACF shows spikes at 1 and 2

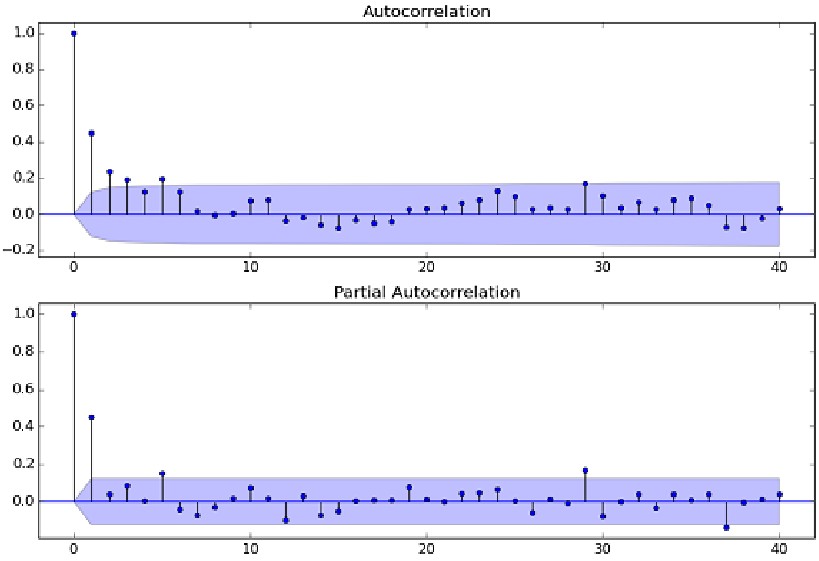

Figure 2: ACF and PAC function plot.

Another way to analyze would be an ARIMA(2,0,1) as we see two significant spikes in PACF and one significant spike in ACF (after which the values die out starting from a much lower point (0.4)). Looking at the in-sample-forecast results (using a simple Mean Absolute Percentage Error) ARIMA (2,0,1) delivers much better results than ARIMA (4,0,2). So we use ARIMA (2,0,1).

2.1.3 Data validation

After fitting the model, the research checked weather the model is appropriate. From Residual analysis, a sequence plot is run to show that the residuals have no constant location and scale. The research preformed a lagged plot to show that the residuals are not auto correlated at lag one. Finally, an auto correlation of the residuals is perform to show that all sample auto correlation falls inside the 95 *per cent confidence interval*

2.1.4 Testing for auto correlation at lag 1

A simple graphical approach is the lagged scatter plot, but the approach is cumbersome when there are many scatter plots to be examined in covering the possibility of relationship at higher lags.

X_{t-1}	*154.98*	*161.671*	*159.215*	*149.86*	*155.08*	*151.00*	*145.43*
X_t	*161.671*	*159.2157*	*149.86*	*155.088*	*151.003*	*145.436*	*141.13*
X_{t-1}	*141.13*	*140.67*	*156.61*	*138.916*	*135.69*	*105.82*	*104.055*
X_t	*140.67*	*156.612*	*138.9163*	*135.699*	*105.824*	*104.0553*	*107.78*
X_{t-1}	*107.7807*	*119.9938*	*126.3438*	*124.121*	*122.6608*	*120.65*	*1220.24*
X_t	*119.99*	*126.3438*	*124.121*	*122.660*	*120.65*	*120.2478*	*109.1142*

.

From the table we plot a graph of the series X_t against its lag X_{t-1}

In our case, the series are auto correlated and therefore the lags are interdependent.

2.1.5 Testing for auto correlation Function of residuals

The auto correlation function shows that all lags fall inside 95 per cent confidence interval, an indication that the residuals are random.
There is no indication of significant autocorrelation in the residuals as confirmed by the Ljung-Box test. The Ljung-Box statistic is 19.8 based on 20 lags, which is not significant (p = 0.65) because the quantile corresponding to the 95th percentile of a chi-squared distribution with 16 degrees freedom is 35.17. The Ljung-Box test is valid under these conditions of non-normality, although for stronger non-normality, the Ljung-Box test is not robust and tends to reject the null hypothesis of no autocorrelation too quickly.

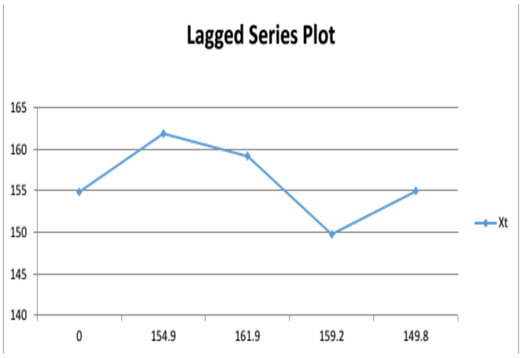

Figure 3: x_t against x_{t-1}

3 Results and Discussion

Data that were collected from Kenya meteorological station in Kisumu is first analysed using twelvemonth moving average and six month central moving average. A time series plot (Fig 1) revealed astationary trend with a slightly increasing seasonal component that has constant mean and variance. Given that it shows stationariness, differencing the time series is not necessary and the integratedpart of the model is taken to be zero.

A plot of Auto correlation and Partial auto correlation factors represented in figure(Fig 2) helps us to identify pattern in the data that is stationary in both mean and variance. The parameters ofthe model are estimated using the mean absolute percentage error that gives an Auto Regressive(AR)term of order(p term) and a moving average term of order (q term) with a value of one. This isARIMA (2,0,1) model.

The data is then subjected for validation . A plot of values of random variable residuals, X against its lag at $X-1$ *as shown in (Fig. 3), shows that most of the spikes fall within the significant line hence there is no correlation within the residuals. This indicates that the residuals are independentlyand identically distributed normal random variables.*

More checks on the residuals using Auto-correlation Function(ACF) shown in (Fig 3) and Partial Auto Correlation Function(PACF) shown in figure (Fig 3*).*

Ljung-Box statistics at lag 20 was found to be 19.8 with a p-value of 0.65, a value that is not significant and lies within the confidence interval. The result shows that the residuals are independent. The forecast values were superimposed on the actual values as shown in (Fig 5) with the aim ofdetermining and comparing the level of accuracy between actual and predicted values including afour years prediction of rainfall from 2010 to 2014 obtained from the model. The $R^2 = 0:90846$ *(a*

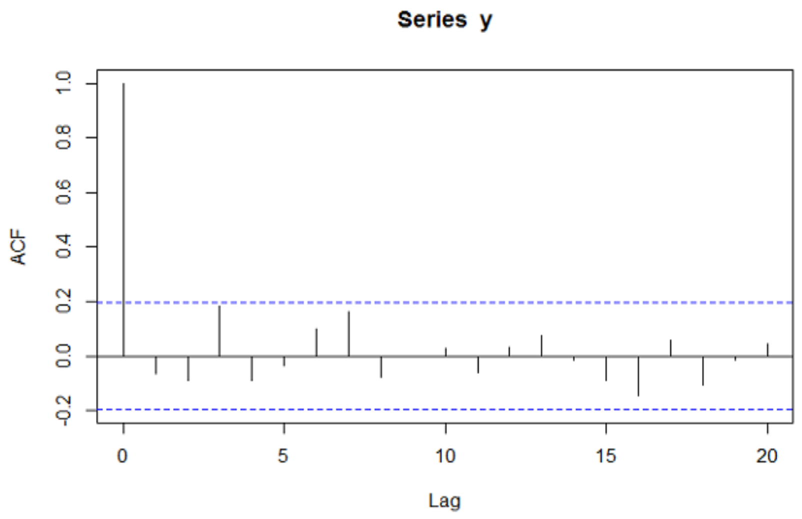

Figure 4: Plot of ACF.

high value), implied that the fitted values are closer to the actual ones. This confirms the ARIMA model (2,0,1) to be accurate and suitable in forecasting rainfall data around Lake Victoria Basin.

4 Conclusions

This study has accurately determined and analysed trends and rainfall patterns around the Lake Victoria region from year's 2007 to 2014 by fitting ARIMA (2,0,1) model from the data set. The study observed the trends and generalized them for use in future forecasting.

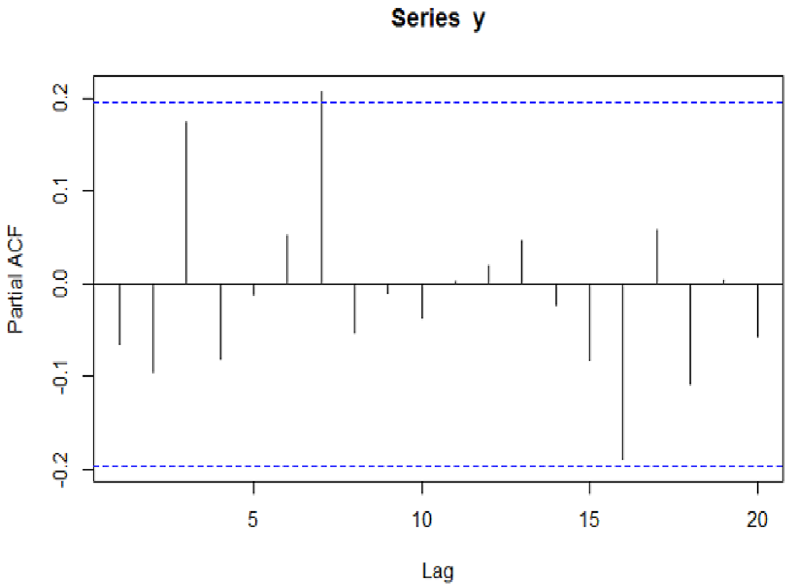

Figure 5: Plot of PACF

References

[1] *Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015).* Time series analysis: forecasting and control. *John Wiley & Sons.*

[2] *Easterling D, Evans J., (2000)*Observed variability and trends in climate events:a brief review. *Bulleting of the American meteoro- logical society, pages 417-425.*

[3] *Lallanilla, M. (2015).*Greenhouse gas emissions: Causes & sources. *Livescience website. Online available at: http://www. livescience. com/37821-greenhouse-gases. html (Accessed on 24, Januray, 2017).*

[4] *Le, T. H., Somerville, R., Cubash, U., Ding, Y., Mauritzen, C., Mokssit, A., ... & Prather, M. (2007). Historical overview of climate change science.*

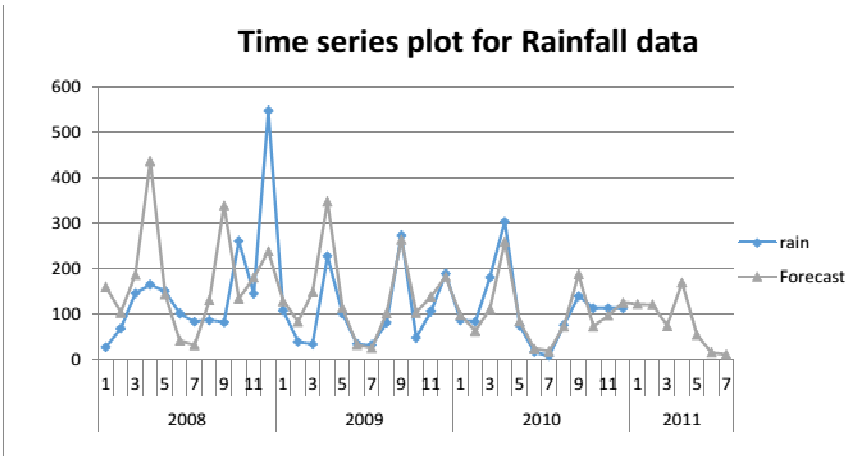

Figure 6: Time series plot of historical and predicted.