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

BY

ROBERT ORYEMA

A RESEARCH PROJECT SUBMITTED IN PARTIAL FULLFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF SCIENCE IN APPLIED STATISTICS

SCHOOL OF MATHEMATICS, STATISTICS AND ACTUARIAL SCIENCE

MASENO UNIVERSITY

DECLARATION

This research project is my own work and has not been presented for a degree award in any other institution.

ROBERT ORYEMA

PG/MAT/0010/2012

Signature ..... Date .....

This research project report has been submitted for examination with my approval as the university supervisor.

Prof. Fredrick Onyango

Department of Statistics and Actuarial Science

Maseno University

Signature ..... Date .....

ACKNOWLEDGEMENT

I give my special thanks and appreciation to my supervisor Prof. Fredrick Onyango for the close guidance and expertise assistance that enabled me to come up with this project work.

I also want to thank my wife Mercy Amenga and daughter Joy Grace for their cooperation during the periods of my studies.

Dedication

I dedicate this project to the almighty God, my wife Mercy Amenga and daughter Joy Grace, my brother Seda Protus, my sister Judith Onyango and my mother Grace Oriema.

Abstract

Economic activities in Lake Victoria Basin such as agriculture, fishing, mining and transportation depend's heavily on the climatic conditions of the Lake and its Basin. Global climatic change caused by Greenhouse Gas emission (GHG) has resulted in a disruptive and erratic weather pattern for these economic activities. This unpredictable weather variations is responsible for loss of life and destruction of property. The primary cause of this negative impact is lack of a reliable information addressing climatic variation within the region. The main objective of this project was to identify a suitable time series model that can be used in predicting, forecasting and analyzing weather variations.The Box jenskin methodology is used to build ARIMA model for rainfall and temperature. Data obtained from the Kenya Meteorological Department's Kisumu, Lwak, and Migori for the years 2007 to 2014 produced An ARIMA (2,0,1) model for rainfall is using R package. Data for the years 2011 to 2013 were estimated using values from the years 2008 to 2010 and the relationship showed a strong positive relationship indicating a high accuracy level on predictability by the model.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

1.1 Background Information

Global temperature increase, known as global warming, is caused by a high concentration of carbon and fluorine related gases in the atmosphere(GHG) as a result of green house gas emission. The gases have a profound effect on environment and affects climatic distribution patterns resulting in global and regional climate changes(Marc,2015). 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(Treut et al.,2007). It is projected that climate change will increase and have an adverse effect on humanity and nature(Easterling et al., 2000).

The environment that supports Lake Victorian basin ecosystem is becoming fragile from such weather changes which take the form of extreme and erratic temperature and rainfall occurrences. Such climatic fluctuations has a significant negative impact on socioeconomic activities along Lake Victoria Basin like crop production, transportation, mining and fish harvesting.

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 inaccurate forecasting include the impending el Nino in Kenya in the months of December, 2014 to February, 2015 and the month of 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 and temperature data for 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(1970). 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. They may possibly include autoregressive terms,moving average terms and differencing operation. ARIMA stands for a series which needs differencing to be made stationary. Lags of the stationarized series are called auto-regressive(AR) terms while lags of the forecast errors are called moving average(MA) terms.

Change in climate may potentially alter the frequency, quantity, lo cation and duration of hydrological regimes. Changed hydrological regimes significantly affects planning and design of water resource structures both in

1.1.1 Definitions

1. **Greenhouse effect**; This is an increase in the earth's atmospheric temperature caused by a high concentration of carbon and fluorine related gases in the atmosphere that trap radiation energy coming from the sun. The presence of significant amount of carbon related gases is caused by the burning of large amount of fossil fuel and forests.

2. **Global Climatic Model**(GCM); This are mathematical models that are applied on climatic condition through computer simulation and then used to predict future climatic changes

3. **Climate forcing**; Different factors that affects the earth climate by shifting the energy balance.

4. **Mesoscale**; The study of weather on the medium scale between ten kilometers to one thousand kilometers.

5. **Inter Tropical Convergence Zone** (ITCZ) is the area around the equator where the North East and South East trade winds come together forming a

higher concentration of rainy cloud

6. **Time series;**Describe's cycles that are observed to occur sequentially over a period of time. The cycles in a time series are composed of trend, seasonal effects, cyclic effects and random effect components.

1.2 Statement of the Problem

Due to changes in weather pattern which has become erratic and unpredictable, farmers practising rain fed agriculture and people involved in other economic activities that depend on the Lake Victoria Basin environment are greatly affected by these climatic changes. These include severe prolonged droughts or sudden flash flooding. Lack of precise guiding information that can enable them manage the effects caused by such changes leads to loss of life and property, a suppressed growth in the region's economy and a rise in poverty level.Lack of a good model that can best fit rainfall and temperature data of the Lake Victoria Basin that can effectively address the changing climatic conditions.

1.3 Objectives of the Study

1.3.1 General Objective

The general objective of the study is to identify a suitable model that can be used in predicting rainfall and temperature along the Lake Victoria Basin

1.3.2 Specific Objective

1. To determine a first gauss at rainfall and temperature over time along Lake Victoria basin and establish a rainfall pattern in Lake Victoria Basin.

2. To determine the parameters of the ARIMA models for the historical rainfall and temperature data for use in weather forecasting.

3. To test the appropriateness of ARIMA model fitted to the Lake Victoria basin 2007-2014 rainfall data.

4. To forecast future rainfall in the Basin.

1.4 Significance of the Study

The study has developed a suitable model (ARIMA) that is precise and applies local regional data which can be used in weather forecasting by the meteorological departments in addressing changing weather pattern.

1.5 Justification of the Study

The model can help in reducing the uncertainty associated with erratic rainfall and temperature distribution such as destruction of property and loss of life from flash flood in the Lake Victorian Basin. Timely dissemination of vital and precise information on weather obtained from the model and given to miners, fishermen and ferry users can reduce weather related tragedies. This is important since about three thousand people die from weather related accidents on Lake Victoria every year, (Tord,2014).

Precise weather forecasting from the model will in future help manage starvation and death from hunger experienced during periods of prolonged drought by enabling farmers, governmental agencies and any other relevant institutions to stock enough food during bumper harvesting season that can letter be used during the dry spell.

The study helps health workers and medical practitioners to be better prepared against outbreaks of vector and water borne disease during warm temperature months from which cholera, bilharzia, malaria and dengue micro organism breeds.

CHAPTER 2

LITERATURE REVIEW

2.1 Seasonal trends in rainfall pattern

Global warming phenomenal has resulted in annual temperature increase of $0 : 6^0$ centigrade(IPCC,2001). The increase has affected global hydrodynamic large scale cycles and individual, different local regions of earth's surface. The local cycles are directly interacted with the global climate with a direct influence on local cycles from the global cycles. Climatic changes with subsequent change in hydrodynamic cycle has resulted in shifting of rainfall and temperature distribution pattern in intensity, duration and frequency. The impact of global climate on local weather was studied(Agwata, 1992)in a research project on Lake Victoria level and global climate change. The results of the finding using Pearson correlation, cross spectral analysis and chi-square correlation, a conclusion was reached that there was a strong link between global climatic cycle and local climatic cycle along Lake Victoria Basin, an attribute to changing of Lake Victoria Basin's rainfall and temperature distributions from the effects of Global Green House Gas consecration.

Lake Victoria and the basin surrounding it significantly depends on, among others, global Inter Tropical Converges zone I.T.C.Z(Yi song,2004.) climatic condition, El nino/southern oscillation(ENSO) and Indian ocean zone's temperature gradient(Black, 2003). Strong interface between Lake Victoria Basin's climate and Inter Tropical Convergence Zone makes the global climatic change to influence hydrodynamics and thermodynamics pattern of the Lake Victorian Basin.

Inter Tropical Convergence Zone crosses East Africa twice every year, once during March-April-May and during October-November-December months of the year. This incursion and retreat of ITCZ is responsible for the two main rainfall season, long rains

in February, march and April and short rains in October, November and December. It is also responsible for the dry seasons of the region in the months of June, July and August. Rainfall around Lake Victoria Basin is characterized with strong winds, heavy downpour accompanied with thunderstorm resulting in loss of life and property.

Rainfall distribution pattern of Lake Victoria Basin depends to a large extent on precipitation from the lake which accounts for 83 per cent of the lake's water source besides the inlet rivers and 76 per cent of water lost from the lake. Precipitation is known to occur mostly between mid- night and early morning hours when there is a strong land breeze due to a warmer temperature of the lake than the adjacent land mass. The highest rainfall is always recorded on the western sector of the lake with a precipitation of 219.4 mm and 205.3 mm for the eastern sector. The floods in 1997-98 affected 900 000 and caused losses equivalent to about 11 per cent of G.D.P.

2.2 Seasonal trends in temperature

Thermodynamics effect on weather pattern including rainfall distribution was noted to be significant by (Yi, 2004) and informed the decision to have a coupled approach in the study of weather pattern along the Lake Victoria Basin. The change in climate due to green house gas concentration in East Africa was projected to be warmer and wetter as global temperature increases, (John, 1999). Global temperature increase at 0.6 degrees centigrade has had a profound effect on regional temperature and other related dynamics. In Africa, it results in the spread of the Sahara desert annually and the disappearance of savannah vegetation, (journal of science, 2014).

The shallowness of Lake Victoria and its large surface area compared to its volume makes it vulnerable to climatic changes.Lake Victoria surface temperature is approximately 240 centigrade. Atmospheric temperature of Lake Victoria Basin is between 21:90 centigrade and 24:30 centigrade. Surface temperature of the lake is known to influence both rainfall and temperature distribution along its basin(xi, 2015)hence influencing the occurrence of flush floods and prolonged drought.

In Kenya, La Nina of 1998-2000 affected 23 million Kenyans and caused losses equiv-

alent to 16 per cent of G.D.P mainly in hydro electric power generation, crop failure, death of livestock and health complication. It has been discovered that environmental warming has a direct effect on disease infection and transmission like malaria and dengue fever which is caused to extend to higher altitudes consequently increasing the infection risk along the Lake Victoria Basin,(Karanja,2006). Malaria is reported to be one of the most climate sensitive vector borne disease(Githeko,2000). Containment of this kind of weather related disease outbreaks by health and disease control personnel can be boosted by availing information on climate variability and being able to predict these variable for future use by the personnel within a required time period.

2.3 Socio-economic activities

Impact of climatic change on hydrological extremes and water resource on Lake Victoria Basin determines different characteristics and other natural resource within it. Lake Victoria is an economic and ecological resource of value to, not only the people living along its basin but also to those that live along the river Nile basin. Some 3.5 million people in Kenya, Uganda and Tanzania depend on the Lake(Global network for climate, 2014). A study (Nyeko, 2011)has shown that change in climate is a real threat to global society and its environment by the occurrence of extremes of climate such as sever drought or prolonged and intense wet spells resulting into floods with a negative impact on both natural and managed system.

The population of the Lake Victoria Basin is about 30 million people(world bank,2007) in all the countries of Kenya, Uganda,Tanzania and Burundi. The region is described as one of the most densely populated areas of the world with some parts of Kenya having a population of 1200 inhabitants per square kilometer(UNEP,. 2001). The vegetation of the lake is 62 per cent woody and herbaceous while 32 per cent of it is agricultural land (Osienala,2015). Economic activities practised by the inhabitant of the lake are large scale rain fed agricultural activities like the sugar plantations in Chemelil, Muhoroni and sony in Nyanza,Kenya, tourism industry, transportation and fishing. Lake Victoria supports the largest fresh water fishery in the world, with a yield in excess of 500,000 tonnes from over 200,000 fishermen and worth US dollars

600 million. Studies has shown a strong positive correlation between rainfall and thus water level and fish landings(Williams,1972). The Lake Victorian Basin has a catchment of the order of 3 to 4 billion dollars of fish annually,(U.N.E.P,2001).

2.4 Studies on Lake Victorian Basin

Many studies have been done by researchers on Lake Victoria and its Basin. (Semazzi,2011), modeled a fully coupled three dimensional empirical study on the climate of the Lake Victorian Basin using mechanistic numerical simulation based on RegCM3-POM coupled model development to determine forcing mechanism such as orographic forcing mechanism, bathymetry and large scale forcing. The study examined relative influence of different processes on Lake Victorian climate.While showing that Lake Victoria does not generate its own climate and its precipitation is greatly enhanced by Easterly trade wind's moisture, the study used a Global climate approach in analyzing weather pattern along Lake Victorian Basin.

Statistical down scaling model was used by (Nyeko, 2011) by match- ing scales from Global circulation model(GCM) with hydrology at Lake Victoria Basin regional scales using extreme value analysis. The outcome of his study was the impact of global climate had on intensity, distribution and frequency of Lake Victoria Basin.

In integrated, regional mesoscale and basin scale approach to climate assessment model(Mutua, 2013) the study took the changes that occurrences in the atmosphere and translated them to the mesoscale and basin scale climate.The results of the study predicted a wetter East African with dry spells on June-August periods. In his research, he noted high changes during the October-December periods whose variability was closely associated with global weather pattern by effects of anomalies of Indian ocean surface temperature.

Most research done on Lake Victoria Basin by researchers involved down scaling climate data from global scale to the regional scale using global circulation model. The main disadvantage of Global climate model is its inability to resolve features smaller than eighty kilometers by eighty kilometers (climate change information resource,

2001). The model incorporate so many factors including radiation, energy transfer by winds, cloud formation, evaporation and precipitation.Although the models have the advantage of performing multiple simulation and hence increasing the level of accuracy on the Global scale, the model,beside not being able to solve smaller regional scale, the model simplify complex and non-linear processes like the effects of radiation by low level clouds or hydrological processes on the land.

Disadvantages associated with GCM gives credence to use of ARIMA model in a localized environment like Lake Victorian Basin that has a lot of land hydrodynamics and low level clouds often formed from precipitations mostly during prolonged rainy season. Extreme climate change can not be wholly attributed to global warming(Muta, 2013) which gives credence to a non Global climatic model approach such as a regional ARIMA model approach.

2.5 Time series theory

Time series is a set of descrete observations made over a period of time. A time series can be expressed as $Y_t = X(t)$ where Y is a stochastic process and X is a random variable. A Time Series can exhibit four types of components namely; horizontal component, when data values fluctuate around a constant value, Trend component, when there is a long term increase or decrease in the data, Seasonal component when a series is influenced by seasonal factors and recurs on a regular periodic basis and the cyclical/ periodic component when the data exhibits rises and falls around trend levels.

2.5.1 Auto-Regressive Integrated Moving Average model(ARIMA)

Forecasting of rainfall and temperature was started by Walker G.T in his publication " seasonal weather and its prediction" (1933). Mathematical models based on probability concepts are applied as stochastic process and used in evaluating weekly, monthly and annual rainfall and temperature data as time series analysis,and has been found

to be better than other methods like spectral analysis model(Asekere, 2004) and the non parametric model(Mitchel, 1966)for predicting both hydrodynamic and thermodynamics. Time series is a sequence of observation made over a period of time and describes variations of such observation as generated from historically recorded data. Analysis of time series through computations and stochastic modeling can be used in weather forecasting and predictions.

Auto regressive (AR) were first introduced by yule (1926) and later generalized by walker (1933), while moving average (MA) models were introduced by Slutzy,(1973) and then wold (1938) provided theoretical foundation for combined auto regressive moving average time series. Raymond Y.C.(1997) suggested two questions that we should put in our minds when using data to analyze time series, first is weather the data are random and second weather the data has any trends. A random series has a correlation that is close to zero and the ARIMA model is not applicable while a series that is statistically dependent on each other is suitable for ARIMA model. The next step in the process is model identification, parameter estimation and testing for model validity. This research used rainfall data to generate the model suitable for both rainfall and temperature. the approach was developed by momani(2009) in his research time series analysis in Jordan.

2.5.2 Basic Concepts

1. A time series is said to be a strictly stationary if the joint probability distribution of the process does not change when shifted in time. Let X_t be a stochastic process, if$X_{t_1}, X_{t_2}X_{t_3}.....X_{t_n}$ is the same as $X_{t_{h+1}}, X_{t_{h+2}}X_{t_{h+3}}X_{t_{n+h}}$ $\forall t_i \in R$. In a stationary series, the mean and variance do not change with time. It has no periodic variations and has no trend with its autocorrelation being constant.

   A time series is called 2nd order stationary or weakly stationary if it has a constant mean and its auto covariance function is independent of time but depends only on the distance between the variables,its mean is finite,i.e if

   $E(X) = \mu < \infty \ \forall t \in \mathrm{R}$

Where E(X) is the expectation of the X

then taking X at times t asX_t it will follow that

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

applying the covariance equation for two variables XY given by

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

$X_t$andX_{t+h} we obtain

$$E(X_t - \mu_t)E(X_{t+h} - \mu_{t+h}) = \sigma(h) \tag{2.1}$$

where$\sigma(h)$ represent the auto covariance at lag h

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 and of X_t andX_{t+h} is given by$\sigma(h)$

   and their respective variances as $V(X_t)$ and $V(X_{t+h})$ respectively

   Then

$$\rho(h) = \frac{\sigma(h)}{\sqrt{V(X_t)V(X_{t+h})}} \tag{2.2}$$

   $\rho(h)$ represents the auto correlation function (ACF) of a time series at a time lag of h between the variablesX_t andX_{t+h} and varies between -1 and +1

3. **Partial autocorrelation function** (PACF) are a measure of correlation between variables$X_t and X_{t+h}$ where there is a large set of lags between them making the auto correlation between them difficult to establish. partial auto correlation function gives the partial correlation within its lagged values handling shorter lag values. the PACF is used in data analysis to identify the extent of a lag in ACF.

   IfΦ_{hh} represent the coefficient of partial regression of the r^{th} order auto regression,

   Then

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

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

$\rho(h) = \Phi_{h1}\rho(j-1) + \Phi_{h+2}\rho(j-2) + ...\Phi_{hh}\rho(j-h)$

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 given by

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

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

$$X_t = \sum_{j=0}^{q} \theta_j e_{t-j} \tag{2.3}$$

θ_1, θ_2, θ_3, ...θ_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

$$var(X_t) = \sigma^2 \sum \theta_j^2 \tag{2.4}$$

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

$$\sigma(h) = \sigma^2 \sum \theta_j \theta_{j+h} \tag{2.5}$$

The auto correlation coefficient for an MA process is

$$\rho(h) = \frac{\sigma^{2\sum \theta_j \theta_{j+h}}}{\sum \theta_j^2} \tag{2.6}$$

5. **Auto regressive process**.

   A time series described by the process

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

   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 $\phi(B) = 0$ lie outside the unit circle in the complex plane where B is a backward shift operator such that

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

   $B^2 X_t = B(BX_t) = X_{t-2}$ etc The AR Model and the AR polynomial

   $\Phi(B) = 1 - \phi_1 B + ...\phi_p B^p$

   For AR(1) we have

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

   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.** This is the combining of autoregressive process (AR(p)) and moving average process (MA(q)) from a stochastic model in order to produce an autoregressive moving average process (ARMA) model that can represent a stationary time series process. The ARMA model is

represented as

$\phi_1 X_1 + \phi_2 X_{t-2} + ...\phi_p X_{t-p} = \theta_1 e_1 + \theta_2 e_{t-2} + ...\theta_q e_{t-q}$

Using a backward shift operator

$$\phi(B)X_t = \theta(B)X_t \tag{2.7}$$

Where

$\phi(B) = 1 - \phi_1 B + \phi_2 B^2 + ...\phi_p B^P$

and

$\theta(B) = 1 - \theta_1 B + \theta_2 B^2 + ...\theta_q B^q$ This equations are polynomials of degree p and q forming an ARMA p,q model. 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 model, the differencing method is used on the data such that $\nabla X_t = X_t - X_{t-1}$

   representing a first order differencing

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

   representing a second order differencing

8. **Auto regressive integrated moving average process.** If a time series has of ARMA model has a trend that is not stationary, it is then integrated by differencing to become stationary.

   If ∇^d represent d times differencing to produce a stationary model , using a backward shift operator function

   $\nabla = (1 - B)$

   this implies $\phi(B)\nabla X_t = \theta(B)e_t$

   can be represented as

$$\phi(B)(1 - B)^d X_t = \theta(B)e_t \tag{2.8}$$

This is an autoregressive integrated moving average model of order p,d,q.

9. **Ljung-Box test;** A test statistics on the residuals of ARIMA model used to confirm if the data are random or not random in nature.

   generally it test the hypothesis .

   H_0 The data are random

   H_1 The data are not random

   The test statistics is given by

$$Q_{LB} = n(n+2)\sum_{j=1}^{h}\frac{P^2}{n-j} \tag{2.9}$$

   Where n is the sample size, p is the auto correlation at lag j, h is the number of lags.

   The hypothesis of randomness is rejected if

   $Q_{LB} = X^2$,

   X^2 being chi square distribution.

CHAPTER 3

ARIMA MODEL

Data collected from kenya meteorological department as secondary data is analyzed using the Box jenskin procedure of

1. model identification

2. parameter estimation

3. data validation.

3.1 results

3.2 Identification of the model

At this stage we try to smooth data using moving average(MA) and central moving average(CMA) process. Smoothing is always done to help us better see patterns and trends in time series by removing irregular roughness and have a clear signal. A 12 period moving average is used to center the 12 months of the year and 6 months central moving average is used as the average of the 12 months. The central moving average helps us to see the trend from the data. The table below represent 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	101.092		6		
7	2008	7	83.312		7		
8	2008	8	86.36		8		
9	2008	9	81.534		9		
1	0 2008	10	260.096		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	127.0786

Table 3.1: Rainfall Moving Average and Central Moving Average

Time	year	month	Rain	year	month	MA(12)	CMA(6)
20	2009	8	80.772		8	156.6122	124.1244
21	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	83.312	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

Table 3.2: Rainfall Moving Average and Central Moving Average

3.2.1 Time series plot of rainfall data

The time series plot for rainfall with the moving average plot shows seasonal behavior of the of the data weather the seasons are increasing, decreasing or remain constant as time increase from which we are able to choose between multiplicative and additive methods. The pot also shows us the general trend of the series in this case a non increasing trend.

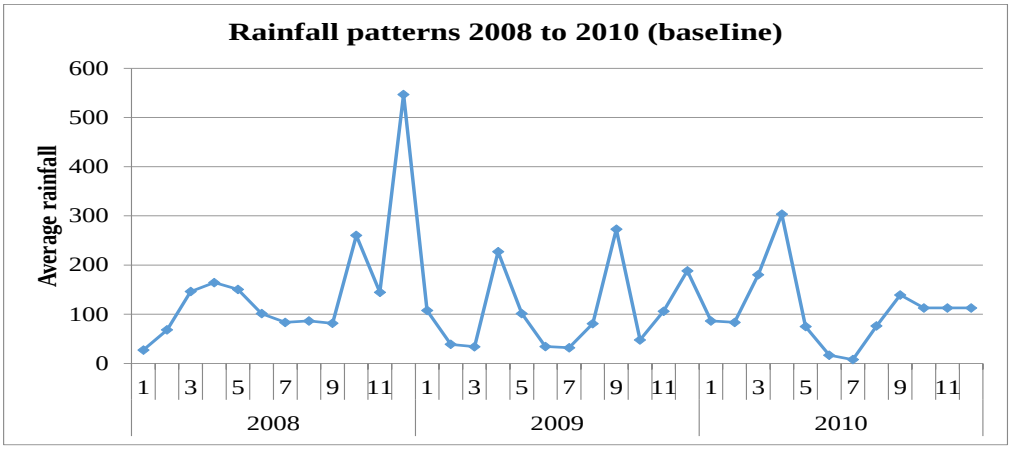

Figure 3.1: Time series plot of rainfall data

3.3 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 after considering other diagnostic methods. Having that in mind, we go ahead and say that 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. Another way to

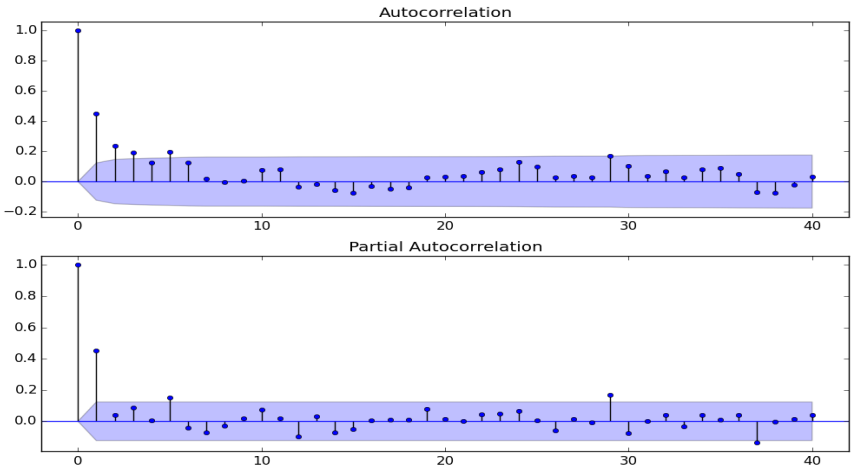

Figure 3.2: ACF and PAC Function plot

analyze would be an ARIMA(2,0,1) as we see two significant spikes in my PACF and one significant spike in my 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).

3.4 Data validation

After fitting the model, we check weather the model is appropriate. using residual analysis, we do run sequence plot to show that the residuals do not have a constant location and scale. We also preform a lagged plot to show that the residuals are not auto correlated at lag 1. Finally, an auto correlation of the residuals is perform to show all sample auto correlation fall inside the 95 per cent confidence interval

3.4.1 Testing for auto correlation at lag 1

A simple graphical approach is always the lagged scatter plot, but this approach is always cumbersome when many scatter plots are to be examined to cover 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	105.105.824	104.0553	107.78
X_{t-1}	107.7807	119.9938	126.3438	124.121	122.6608	120.65	120.24
X_t	119.99	126.3438	124.121	122.660	120.65	120.2478	109.1142

Table 3.3: A table ofX_t against X_{t-1}

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

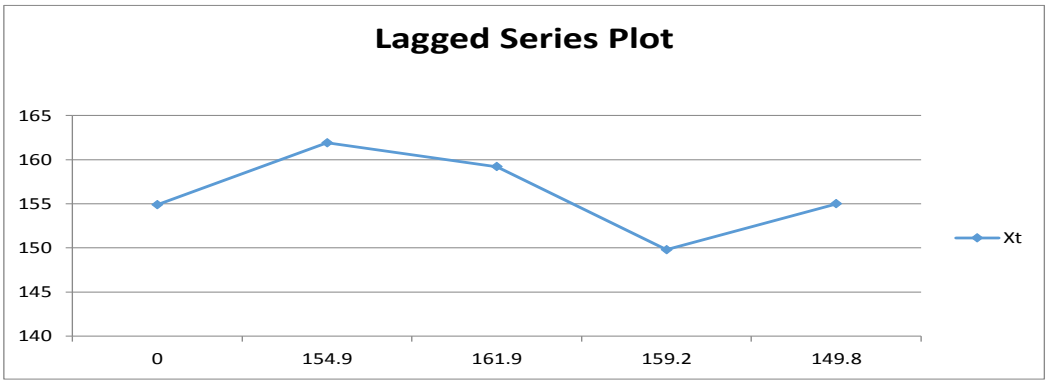

Figure 3.3: X_t agaist X_{t-1}

In our case, the series are auto correlated and therefore the lags are interde-

pendent

3.4.2 Testing for auto correlation Function of residuals

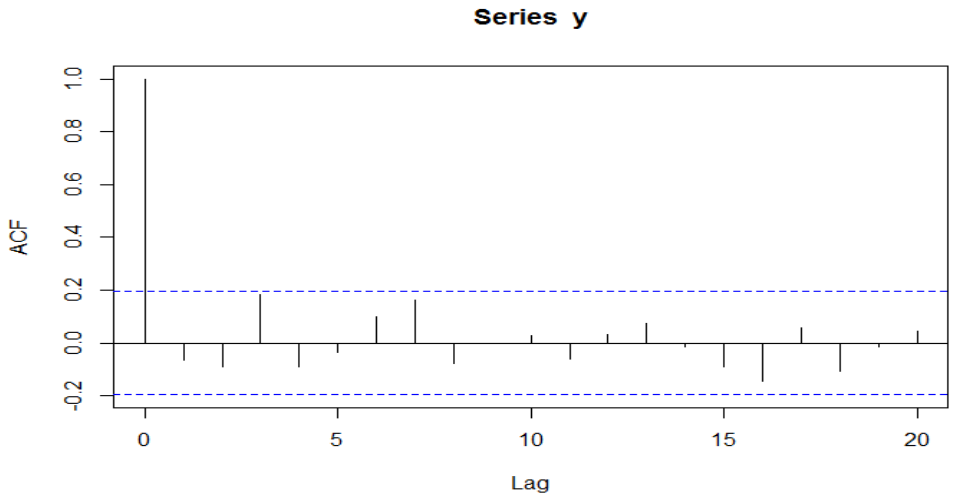

Figure 3.4: plot of ACF

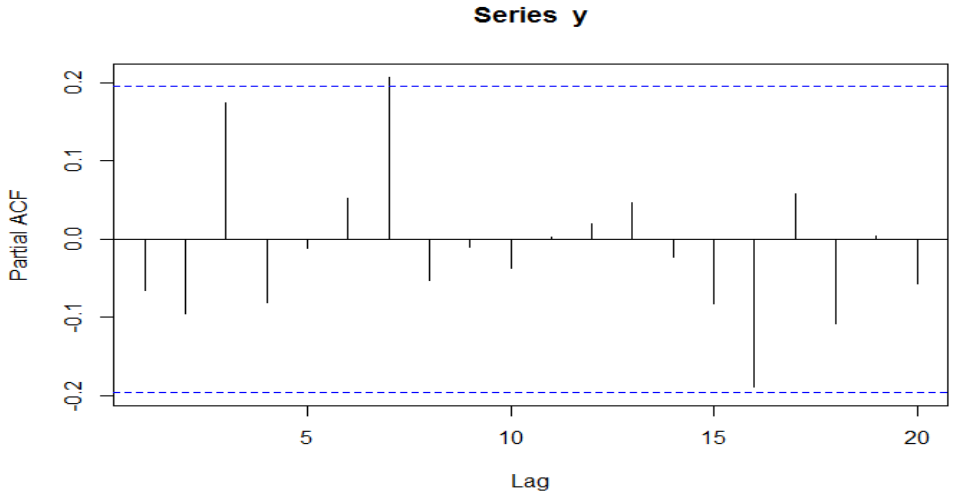

Figure 3.5: plot of PACF

The auto correlation function shows that all lags fall inside 95 per cent confidence interval indicating the residuals to be random.

There is no indication of significant autocorrelation in the residuals, which was confirmed by the Ljung-Box test. The Ljung-Box statistic was 19.8 based on 20 lags, which was 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.

CHAPTER 4

RESULTS AND DISCUSSION

Data that were collected from kenya meteorological station in kisumu was first analysed using a 12 month moving average and a 6 month central moving average. A time series plot (Fig 3.1) revealed a stationary trend with slightly increasing seasonal component that was constant in mean and variance. Given that it was showed stationarity, differencing the time series was not necessary and the integrated part of the model was taken to be 0.

A plot of Auto correlation and Partial auto correlation factors(Fig 3.2) helps us to identify pattern in the data which is stationary in both mean and variance.possible model. The parameters of the model are then estimated using the mean absolute percentage error giving us Auto Regressive term of order(p term) 2 and a moving average term (q term) value of 1. This gives us ARIMA 2,0,1 model.

The data was then subjected to validation . A plot of the values of random variable residuals of X against its lag at X-1 (Fig. 3.3)showed most spikes falling within the significant line hence no correlation within the residuals indicating that the residuals are independent and identical normally distributed random variables. Further checks were done on the residuals using auto correlation function (Fig 3.4) and partial auto correlation function (Fig 3.5). Ljung-Box statistics at lag of 20 was 19.8 with a p value of 0.65 which was not significant an indicated that it lies within the confidence interval hence the residuals are independent. The forecast values were superimposed on the actual values (Fig 5.1) with a view to compare and determine the level of accuracy between the actual values and the predicted values and a four years prediction of rainfall from 2010 to 2014 obtained from the model. The $R^2 = 0{:}90846$ (a high value), implied that the fitted values are closer to the actual ones.This confirms the accuracy of the model and that it can be used in weather forecasting in

the Lake Victoria Basin. The result shows the ARIMA model (2,0,1) to be a suitable model for the this rainfall data.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusion

From this study we have accurately determined and studied the trends and rainfall patterns in the Lake Victoria region from 2007-2014.We have therefore fitted the most appropriate ARIMA (2,0,1) model from these data and observed the trends which can be generalized for the future years and can be used in forecasting.

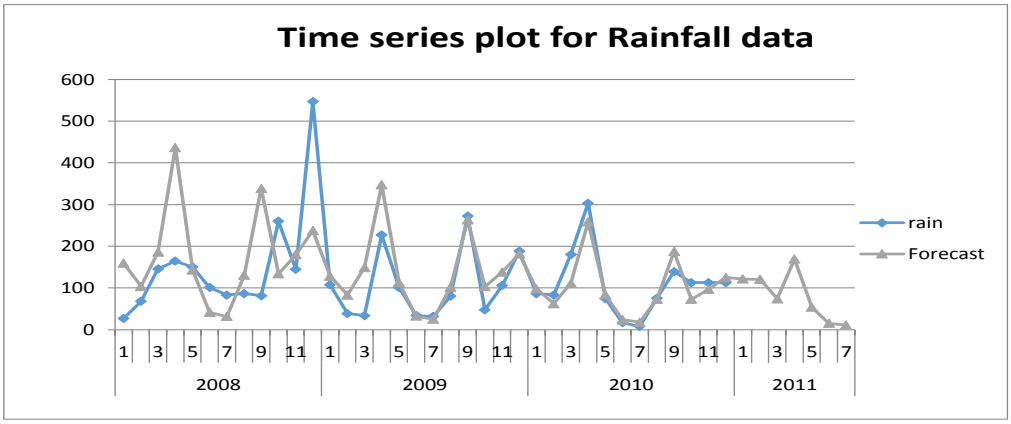

Figure 5.1: Time series plot of data historical and predicted

5.2 Recommendation

We therefore encourage the people living around Lake Victoria region to adopt this model especially those practicing agriculture and any other economic activity in relation to weather determination to enable them manage their activities appropriately and avoid the negative impacts associated with weather.The model is also flexible and can be adjusted from time to time incase there are adverse deviations from the normals as contained in the model.

REFERENCES

[1] **A. Mano, K. Udo,** (2014). Down scaling weather forecast output using ANN for flood prediction, *journal of applied mathematics*, pp 1-14.

[2] **B. George, Jenkins** (1970). *Time series analysis*. Forecasting and control.

[3] **Climate change information source,** (2001). What is global climate model,http://ccir.ciesin.columbia.edu/nyc.

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

[5] **D. M. Sambwa, J. Karanja** M.,(2006). *Environment for development.*An ecosystem assessment of Lake Victorian Basin Environmental and socio economic status,trend and human vulnerability. Health, nutri- tion in the Lake Victorian Basin,p 5.

[6] **E. Black** (2003). *D*ipole model index,Mon. Wea. Rev.

[7] **F. Mutua** (2013). *Integrated climate change impact assessment and extreme event forecasting on Lake Victoria basin*, page 17.

[8] **F. Semazzi, H. Fredrick.** (2011) A modelling and Empirical study of the climate of Eastern Africa, https://ncsu.pure.elsvier.com.

[9] **Global network for climate** (2014). Lake victoria storm warning system, Blending technology with culture. www.gfcs.climate.org,.p 1.

[10] **G.T. Walker** (1938). *Seasonal weather and its prediction*,British association of sciences, Vol 103,pp 25-44

[11] **H. Asekere** (2004). Statistical analysis on annual temperature changes.

[12] **I.P.P.C** (2001). Assesing impact of climate change on Lake Victoria, wedc.iboro.ac.uk/resource.

[13] **J. Agwata.** (1992). The response of lake Victoria level to regional and global climate change, FM.

[14] **J. Dewa.** (2006). Lake Victorian Basin Environmental outlook,*environment for development,UNEP* p 1.

[15] **J. M. Mitchel.** (2014). Climate change. *Technical notes no. 195.* pp 2-5,60

[16] **journal of science.** (2014). *Global warming and sub tropical bound- aries though spread of sahara desert. vol 312* pp 1179

[17] **Marc, Lallanilla.** (2015). Greenhouse Gas Emissions: causes and sources, live science,m.livescience.com/37821-greenhouse.

[18] **Momani, P.E, Nail, M.** (2009). Time series analysis model for rainfall data in Jordan: case study for using Time series Analysis, *American journal of enviromental science* 5(5): 599-604

[19] **P. Nyeko.** (2011). Climate change impact on hydrological extremes and water resources in Lake Victorian catchment, upper Nile basin, http://lirias.kuleuven.be/handle

[20] **P. Wilmott, S. Howinson and J. Dewyne.** (2001). *Lake Victorian Basin outlook.* Demographic and ethnic composition, pp 3.

[21] **R. Anya, F. Semazzi, L. Xie.** (2006). Simulated physical mechanism associated with climate variability over Lake Victorian basin in East Africa. *Monthly weather review*, pages 3588-3609. (2006).

[22] **R. Tord** (2014). Improving weather forecasting in Africa,UIB Global

[23] **Treut, L. Somerville, R. Cubasch, U. Ding, Y. Maurtizen, C. Mokksit, A. Peterson, T. Pranther, M. QIN, D. Manning, M. Chen, Z. Marquis, M. Averty, K.B. Tinor.** Historical overview of climate change science.

[24] **World Bank.** (2007). Environmental policy on Lake Victoria Basin, *World bank journal* pp 11

[25] **Xian, Sun.** (2015) *Lake surface temperature*. The relationship of Lake surface temperature on spatial distribution and intensity of precipitation over the Lake Victorian Basin, mon wea. Rev. 143, 1179-1192. 13.

[26] **Y. C. Raymond.** (1997). application of ARIMA model to real estate.

[27] **Y. Song.** (2004) A coupled regional climate model for the Lake Victorian basin of east africa. *international journal of climatology*, pp 58-75.