APPLICATION OF SEASONAL AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (SARIMA) TO MODEL AND FORECAST WATER DEMAND IN KISUMU CITY, KENYA

by

Nyabwanga, Robert Nyamao

A Project Report Submitted in Partial Fulfillment

of the Requirements for the Degree of

Master of Science in Applied Statistics

School of Mathematics, Statistics and Actuarial Science

MASENO UNIVERSITY

Abstract

Increased population of Kisumu City over the past years has resulted into high demand and competition for water and related facilities. This is evident in the persistent water scarcity within the City, use of poor quality water by the residents and inequitable water distribution. The current net water supply capacity of Kisumu City Water Supply System is 5,400m^3/day against a demand of 27,000 m^3/day. Effective planning and management of the City's water resources is therefore critical in providing reliable forecasts. Models developed for such forecasts ought to take into account the non stationary and seasonality behaviours exhibited by residential water demand data. Research on residential water demand in the Kenyan context have used Ordinary Least Squares, a methodology that does not model the seasonality aspect. In the $SARIMA(p,d,q)(P,D,Q)_{12}$ which is expressed as $\phi_p(B)\Phi_P(B^S)\nabla^d\nabla_S^D X_t = \theta_q(B)\Theta_Q(B^S)e_t$, B^S allows for the modelling of the seasonal behaviour in the data. However, the application of SARIMA to model and forecast residential water consumption in the Kenyan Context is scanty. The study therefore sought to propose a SARIMA model for forecasting residential water demand using secondary monthly water consumption data obtained from KIWASCO for the years 2004 to 2013. Preliminary investigation of the data showed that the data followed a 3-parameter log-normal distribution. Therefore, using logarithm values of the data, the study established by both OLS and Kendall's tau test that the residential water demand for Kisumu City had a significant increasing trend. The KPSS and ADF tests revealed that the data had unit roots which were however removed by first difference. The Data was then fitted to a SARIMA model and the parameters of the model were estimated using Maximum Likelihood Method. $SARIMA(1,1,1)(0,1,1)_{12}$ had the least BIC and AIC values of 2205.273 and 2197.282 respectively and was identified as the better fitting model. Compared to the OLS model, $SARIMA(1,1,1)(0,1,1)_{12}$ had the least MAPE and RMSE values of 3.59 and 7476.59 respectively implying that it had higher forecasting performance. One year forecasts for 2014 were established together with their CI. The observed values for January and February 2014 were within the Confidence Limits. The study recommends the integration of the model by KIWASCO and other water companies in their design of water demand management policies.

Chapter 1

Introduction

1.1 Background of the Study

The scarcity of water resources is a crucial problem in almost every contemporary society. Even in areas where there is adequate water resources, the problem of scarcity is usually confronted by deterioration of water quality leading to increased costs for the users.The problem manifests itself in increased costs of water use, intensified competition over access to water resources and breakout of diseases due to lack of water. Urbanization, population growth, industrial development of cities and rising living standards have led to a growing trend in the percapita consumption of water. Because of the economic investment needed to develop new water resources, accurate prediction of water demand is very important than ever [8]. At the 2000 Millennium Summit held in New York, member countries of the United Nations unanimously agreed on a set of 8 goals to reduce poverty by 2015: among which is reducing by half the proportion of households that do not have access to safe water [8]. According to Worthington [36], residential water demand covers uses of water by households, both inside and outside the confines of the residence and typically includes washing, cooking, bathing, laundry and gardening and its use is usually shown to be highly sensitive to seasonal fluctuations.

According to UNDP report of 2006 [31], 2040 is a more likely date for this goal to be reached in Africa unless there is accelerated investment in the sector. By estimating the proportion of general population accessibility to piped water at home, the estimate also provides an estimate of the number of people potentially exposed to water-related health

risks.

World Bank Report of 2005[35] observes that poor access to water supply is often a result of poor policies and management practices. However, there is significant disagreement over the approach to addressing the problem. Likewise in the World Bank Report of 2003 [34] it is argued that a first and crucial step towards improving water situation and its management is to treat water as an economic good. The Economist of July 19-25, 2003, argues that the problem above all, is that it has been colossally under-priced and that to meet the target of halving the proportion of people without access to clean water money will play a part. Water is therefore viewed as an economic as well as social good. But greater reliance on pricing and markets are even more crucial

Council of Kisumu under the Companies Act Chapter 486 of the Laws of Kenya, established a water company by the name Kisumu water and Sewerage Company (KIWASCO) in 2001 but became operational in 2003. KIWASCOs mandate is to effectively and efficiently provide adequate water to customers and collect, treat and dispose sewerage in a safe and environment friendly manner [15].

In Kisumu City, the mean household water consumption is 149.50 l per day, resulting in a mean per capita of 32.92 L per day. According to a study by [35], the daily per capita water use in Kenya is 45.2 L. Using the recommended basic water requirement of 50 l/c/d by [34], in the study area, there is a mean daily water per capita shortfall of 17.18 L. Wagah et al.[32]further assert that only 25 percent of the households access the minimum recommended basic water requirement of 50 l/c/d.

Kisumu City residents obtain water from individual connections, yard tap connections, public tap connections, boreholes, springs and water vendors. As of September 2008, KIWASCO had 7,704 domestic water connections and 287 water kiosks [32]. About 52 percent of Kisumu residents used piped water delivered to dwellings or compounds, and 13 percent depended on protected shallow wells/springs or roof catchment [32]. Hence 65 percent of Kisumu residents had access to an improved water source, while 35 percent relied on unimproved water sources, including water vendors, open wells/springs, streams and ponds [32]. However, Most residents in informal settlements only have access to water of poor quality, mainly because their water often comes from shallow wells and

water vendors.

There has been a steady increase in population over the years with no expansion in supply capacity. This high population growth rate for Kisumu City has resulted into increased demand and competition for water and the related facilities causing water scarcity [32]. As a result, the water deficit has continued to grow. The current projected water production is 18,000 m^3, while the present demand is estimated to be 48,000 m^3 [15].This indicates a big short fall which must be met by other sources. Further, Maoulidi [19] projects that in the years 2012 to 2015, water supply needs would be 50,000 m^3 per day.

There is an expected large expansion in the fisheries, and the recently revived Molasses Plant which relies on raw molasses from the sugar factories within the region of Nyanza and Western provinces. This poses a huge and increased demand for water and the related infrastructure. Also, revival of East African Cooperation in which Kisumu is expected to play a leading role as the most central commercial and international trade centre with Uganda, Tanzania, Burundi and Rwanda is expected to lead to increased demand for planned development of housing, commerce and industries which will at the same time lead to an increased demand for water resources. This calls for proper forecasting of such demand to guide in water production and management decisions.

Larger efforts to model the demand for water has applied Ordinary Least Squares methodology to estimate the influence of variables such as water prices, housing density, level of income, household size, rainfall and temperature on residential water use. in a functional notation $WaterDemand = f(prices, housingdensity, levelofincome, household\ size, rainfall, temperature, NumberofWomeninhouseholds)$. As observed by Gupta [9], the OLS methodology seeks to minimize the sum of the squares of the deviations. i.e.$\sum(y - y_c)^2$. Using the OLS procedure, Bithas and Stoforos [1], Yaw [38] showed that that Household size, gender of the household head, increasing income, real GDP, Real water price and trend had a significant influence on residential water demand, findings that were in contrast to findings by Xinming, Dale and Briscoe [37] who established that only the number of women in the households as a proportion of total household size had a significant effect on water demand at 0.05 level of significance.

As observed by Jorge [12] the use of OLS approach to model water demand based of economic and social variables ignores the impact of trend and seasonal variations. This can account for the inconsistencies in the OLS findings. Jorge [12] further contends that water demand is highly dominated by daily,weekly, monthly and yearly seasonal cycles which can best be modelled using univariate time series models. Such methodologies are based on the historical data series and are quite useful for short-term demand forecasting since they accommodate the various periodic and seasonal cycles in the model specifications and forecasts.

Following the Box−Jenkins approach, the study applied SARIMA to model the residential water demand time series data in order to propose the best fitting model and forecast future monthly water demand for Kisumu City. The choice of this type of model was based on the established behaviour of the water demand data. [27] observes that the SARIMA is a significant methodology of modelling data exhibiting seasonal behaviour. Also According to Caldwell [5], this Box-Jenkins methodology is particularly suited for development of models for processes exhibiting strong seasonal behaviour.

Box and Jenkins [3] have developed a practical procedure for choosing an appropriate ARIMA model out of the family of ARIMA models. The ARIMA models are especially suited for short term forecasting because they place more emphasis on the recent past rather than distant past. This emphasis on the recent past means that long-term forecasts from ARIMA models are less reliable than short-term forecasts [25].

1.2 Statement of the Problem

The rate of population growth, migration for work, long periods of warm years and the dust factor have over the years caused an increase in water demand in Kisumu City leading the current net demand of 27000 m^3/day. The Kisumu Water Supply System, managed by KIWASCO, has a current net capacity of 5400 m^3/day which is far much below the demand. For the City's water supply system to efficiently bridge the gap between water supply and water demand there is need to accurately forecast the City's water demand. Empirical models that have been developed to forecast water demand take into account

the effect of earlier events on the demand for water and have employed ordinary least squares (OLS)and generalised least squares (GLS)techniques. These techniques do not take into account the strong seasonal behaviour exhibited by water demand time series data and therefore they may be unable to provide accurate forecasts. For effective water demand management, accurate future forecasts are essential hence the need for statistical models that could provide fairly accurate forecasts. This study therefore sought to build a Seasonal Autoregressive Integrated Moving Average (SARIMA) model using monthly residential water consumption data (2004– 2013). These type of models are known to be very robust and do provide respectable forecast performances beside the fact that they do take into account the seasonality aspect in data.

1.3 Objectives of the study

1.3.1 General Objective

The study sought to develop a univariate SARIMA model for residential water demand for Kisumu city that was to be applied to forecast future water demand in the city.

1.3.2 Specific Objectives

The specific objectives of the study were;

(i) To analyse the trend of residential water demand in Kisumu City for the years 2004 to 2013

(ii) To propose a SARIMA model that can be used to forecast residential water demand in Kisumu City

(iii) To forecast residential water demand for Kisumu city in the twelve months of 2014

1.4 Significance of the study

- This will be an essential component in designing effective water demand management policies that can help and guide decision makers establish strategies, priorities and proper use of water resources in Kenya.
- The results will aid KIWASCO in planning because many important water decisions depend on the anticipated future values of demand rates. Also, the results will allow KIWASCO and other water providers in Kenya to explore realistic decision making scenarios for designing effective water demand management policies.
- It will also be crucial in determining water prices and evaluating water investment projects.

1.5 Basic concepts

The study employed the concept of time series which is a stochastic mechanism that gives rise to observed series useful in predicting future events. The basic concepts and theories related to this are discussed below

1.5.1 Time Series

According to Cryer and Kung-Sik [6], a time series is a chronological sequence of observations on a particular variable. The purpose of time series analysis is to understand the stochastic mechanism that gives rise to an observed series and to predict future events or values of that series. Observations may be made of a continuous time series at regular intervals or be aggregations of discrete events. When describing time series $\{X_t\}$ typically represents the observations made at time t, and it is assumed that observations are made at intervals equally spaced in time.

Forecasts are based upon results on an analysis of past events. The first step in the analysis of historical data is the identification of a pattern that can be used to describe the data. This pattern is extrapolated into the future in order to prepare a forecast. This

basic strategy is applied in most forecasting techniques and rests on the assumptions that the pattern identified will continue into the future.

stationary time series

If X_{t1}, X_{t2}.....X_{tn} are observation at time t_1,t_2.....t_n; then the series will be said to be stationary if:

$$E[X_t] = \mu$$
$$Var(X_t) = \sigma^2 \tag{1.1}$$

and both are Constants

Definition 1:

A time series $\{X_t\}$ is said to be strictly stationary if the joint distribution of X_{t1}, X_{t2}.....X_{tn} is the same as the joint distribution of X_{t1+h}, X_{t2+h}.....X_{tn+h} for all t_1,t_2.....t_n and h being real numbers. In other words, strict stationarity requires that the joint distribution of (X_{t1}, X_{t2}.....X_{tn}) be invariant under time shift [27]

Definition 2:

A time series $\{X_t\}$ is said to be weakly stationary or second order stationary if its mean is constant and its Auto Covariance function is independent of time and only depends on the time lag between the variables [27]. i.e.

$$E(X_t) = \mu \tag{1.2}$$

$$\begin{aligned} Var(X_t) &= E(X_t - \mu)((X_{t+h} - \mu) \\ &= Cov(X_t, X_{t+h}) \\ &= \sigma^2(h) \end{aligned} \tag{1.3}$$

Definition 3:

Let $\{X_t\}$ be a stationary time series. The autocovariance function (ACVF) of $\{X_t\}$ at lag h denoted by $\gamma_x(h)$ is given by:

$$\gamma_x(h) = Cov(X_{t+h}, X_t). \tag{1.4}$$

which is called the lag h autocoviance of $\{X_t\}$. It has two important properties:

(i) $\gamma_0 = Var(X_t)$

(ii) $\gamma_{-h} = \gamma_h$

Definition 4:
The autocorrelation function (ACF) of X_t at lag h is given by:

$$\rho_x(h) = \frac{Cov(X_t, X_{t+h})}{\sqrt{Var(X_t), Var(X_{t+h})}} = \frac{\gamma_x(h)}{\gamma_x(0)} \tag{1.5}$$

From the definition, we have $\rho_0 = 1$, $\rho_h = \rho_{-h}$, and $-1 \leq \rho_h \leq 1$. In addition, a weakly stationary series X_t is not serially correlated if and only if $\rho_h = 0$ for all h > 0.

Time series models which are a class of stochastic process have evolved from a simple process to more sophisticated processes depending on the underlining structure. These are as discussed below:

1.5.2 White Noise or Purely Random Process

A discrete stochastic process is called a white noise if it consists of random variables e_t which are Identically and Independently Distributed with $E(e_t) = 0$ and $Var(e_t) = \sigma^2$ [27]. Further for a white noise process the autocovariance function is given by:

$$\gamma(h) = \begin{cases} \sigma^2 & \text{if } h = 1 \\ 0 & \text{otherwise} \end{cases}$$

and the autocorrelation function for the white noise is given by:

$$\rho(h) = \begin{cases} 1 & \text{if } h = 0 \\ 0 & \text{otherwise} \end{cases}$$

1.5.3 Random Walk Process

Suppose e_t is a discrete purely random variable with mean of 0 and variance, σ^2; a process $\{X_t\}$ is said to a random walk iff:

$$\begin{aligned} X_t &= X_{t-1} + e_t \\ &= \sum_{i=1}^{t} e_i, with X_0 = 0 \end{aligned} \tag{1.6}$$

The mean and variance of the random walk process are given as follows:

$$\begin{aligned} E(X_t) &= E(\sum_{i=1}^{t} e_i) \\ &= \sum_{i=1}^{t} E(e_i) \\ &= t\mu \\ &= 0 \end{aligned} \tag{1.7}$$

and

$$\begin{aligned} Var(X_t) &= Var(\sum_{i=1}^{t} e_i) \\ &= \sum_{i=1}^{t} \sigma^2 \\ &= t\sigma^2 \end{aligned} \tag{1.8}$$

Since $Var(X_t) = t\sigma^2$ is dependent on t, the random walk process is non−stationary

1.5.4 Autoregressive Process – AR(p)

The autoregressive structure is a stochastic process that assumes that current data can be modelled as a weighted summation of previous values plus a random term. The process is regressed on past values of itself and this explains the prefix auto in the regression process. Assume the random term; e_t is purely random with mean zero and Variance

(σ^2); then $\{X_t\}$ as an autoregressive process of order p written as AR (p) is given by:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_3 X_{t-3} \ldots\ldots\ldots\ldots \alpha_p X_{t-p} + e_t \tag{1.9}$$

Using the Backward Shift operator B, the equation above becomes;

$$\begin{aligned} X_t &= \alpha_1 B X_t + \alpha_2 B^2 X_t + \alpha_3 B^3 X_t \ldots\ldots \alpha_p B^p X_t + e_t \\ X_t - \alpha_1 B X_t - \alpha_2 B^2 X_t - \alpha_3 B^3 X_t \ldots\ldots\ldots\ldots - \alpha_p B^p X_t &= e_t \\ (1 - \alpha_1 B - \alpha_2 B^2 - \alpha_3 B^3 \ldots\ldots\ldots\ldots - \alpha_p B^p) X_t &= e_t \\ X_t = \frac{e_t}{1 - \alpha_1 B - \alpha_2 B^2 - \ldots\ldots - \alpha_p B^p} & \end{aligned} \tag{1.10}$$

In particular, the first order Autoregressive process AR(1) provided that $\mid \alpha \mid < 1$ will be written as:

$$\begin{aligned} X_t &= \frac{e_t}{(1 - \alpha B)} \\ &= e_t(1 + \alpha^1 B + \alpha^2 B^2 + \alpha^3 B^3 + \ldots\ldots\ldots\ldots) \\ &= \sum_{i=0}^{\infty} \alpha^i e_{t-i} \end{aligned} \tag{1.11}$$

The Mean of $\{X_t\}$ for the AR(1) process is thus given as follows:

$$\begin{aligned} E(X_t) &= E(\sum_{i=0}^{\infty} \alpha^i e_{t-i}) \\ &= \sum_{i=0}^{\infty} \alpha^i E(e_{t-i}) \\ &= 0 \end{aligned} \tag{1.12}$$

The Variance of X_t for the AR(1) process is given by:

$$\begin{aligned} Var(X_t) &= Var(\sum_{i=0}^{\infty} \alpha^i e_{t-i}) \\ &= \sum_{i=0}^{\infty} \alpha^{2i} Var(e_{t-i}) \\ &= \sum_{i=0}^{\infty} \alpha^{2i} \sigma^2 \end{aligned}$$

$$\begin{aligned} &= \sigma^2 \sum_{i=0}^{\infty} \alpha^{2i} \\ &= \sigma^2(1 + \alpha^2 + \alpha^4 .....) \\ &= \frac{\sigma^2}{1-\alpha^2} \end{aligned} \tag{1.13}$$

The Autocovariance (ACVF) is also given by:

$$\begin{aligned} Cov(X_t, X_{t+h}) &= \gamma(h) = E(\sum_{i=0}^{\infty} \alpha^i e_{t-i})(\sum_{j=0}^{\infty} \alpha^j e_{t+h-j}) \\ &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \alpha^i \alpha^j E(e_{t-i}, e_{t+h-j}) \end{aligned}$$

if we let j=i+h we have;

$$\begin{aligned} \gamma(h) &= \sum_{i=0}^{\infty} \alpha^{2i} \alpha^h E(e_{t-i}, e_{t-1}) \\ &= \sigma^2 \alpha^h \sum_{i=0}^{\infty} \alpha^{2i} \\ &= \sigma^2 \alpha^h [1 + \alpha^2 + \alpha^4 + ...] \\ &= \sigma^2 \alpha^h [\frac{1}{1-\alpha^2}] \\ &= \frac{\sigma^2 \alpha^h}{1-\alpha^2} \\ &= \sigma^2 \alpha^h, if \mid \alpha \mid < 1 \end{aligned} \tag{1.14}$$

Definition 5:Autocorrelation funtion (ACF)

Considering the first order Autoregressive process, the autocorrelation function is given as:

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{\frac{\sigma^2 \alpha^h}{1-\alpha^2}}{\frac{\sigma^2}{1-\alpha^2}} = \alpha^h \tag{1.15}$$

1.5.5 Moving Average-MA(q)

Suppose that e_t is a white noise with mean 0 and variance σ^2, then $\{X_t\}$ is said to be a moving average process with order q, Written as MA(q) if:

$$\begin{aligned} X_t &= \beta_0 e_t + \beta_1 e_{t-1} + \beta_2 e_{t-2} \ldots\ldots\ldots\ldots \beta_q e_{t-q} \\ &= \sum_{i=0}^{q} \beta_i e_{t-i} \end{aligned} \tag{1.16}$$

The process is said to be weakly stationary because the mean is constant and the covariance does not depend on time t but on the time lag between the variables. The mean and variance of the MA process is then given by:

$$\begin{aligned} E(X_t) &= \beta_0 E(e_t) + \beta_1 E(e_{t-1}) + \ldots\ldots\ldots\ldots \beta_q E(e_{t-q}) \\ &= 0 \end{aligned} \tag{1.17}$$

and

$$\begin{aligned} Var(X_t) &= Var(\sum_{i=0}^{q} \beta_i e_{t-i}) \\ &= \sum_{i=0}^{q} \beta_i^2 Var(e_{t-i}) \\ &= \sigma_e^2 \sum_{i=0}^{q} \beta_i^2 \end{aligned} \tag{1.18}$$

The autocovariance($\gamma(h)$) function of the Moving Average process is given by:

$$\gamma(h) = \begin{cases} 0 & \text{if } h > 0 \\ \sigma^2 \sum_{i=0}^{q-h} \beta_i \beta_{i+h} & \text{if } 0 < h < q \\ \gamma(-h) & \text{if } h < 0 \end{cases}$$

The Autocorrelation ($\rho(h)$) function of the Moving Average process is given by:

$$\rho(h) = \begin{cases} 1 & \text{if } h = 0 \\ \frac{\sum_{i=0}^{q-h} \beta_i \beta_{i+h}}{\sum_{i=0}^{q} \beta_i^2} & \text{if } h = \pm 1, \pm 2, ..., \pm q \\ 0 & \text{Otherwise} \end{cases}$$

To ensure that there is a unique MA process, the property of invertibility must be ensured.

Consider a MA(q) and letting $B^i X_t = X_{t-i}$ for all i be a backward shift operator then the MA(q)can be expressed as:

$$\begin{aligned} X_t &= \beta_0 e_t + \beta_1 e_{t-1} + \beta_2 e_{t-2} ..... \beta_q e_{t-q} \\ &= \beta_0 B^0 e_t + \beta_1 B^1 e_t + \beta_2 B^2 e_t ..... \beta_q B^q e_t \\ &= (\beta_0 B^0 + \beta_1 B^1 + \beta_2 B^2 ..... \beta_q B^q) e_t \\ &= \phi(B) e_t \end{aligned} \tag{1.19}$$

where $\phi(B)$ is a polynomial of order q

X_t is invertible if the roots of the equation $\phi(B) = 0$ all lie outside a unit circle.

1.5.6 Autoregressive Moving Average -ARMA(p,q) Process

Basically, an ARMA model combines the ideas of AR and MA models into a compact form so that the number of parameters used is kept small. An ARMA(p,q) is written as:

$$\begin{aligned} X_t &= \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} + e_t + \beta_1 e_{t-1} + \beta_2 e_{t-2} + ... + \beta_q e_{t-q} \\ &= \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=0}^{q} \beta_i e_{t-j} \end{aligned} \tag{1.20}$$

Using the backward shift operator (B), the above equation can be written as:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + ... + \alpha_p X_{t-p} + e_t + \beta_1 e_{t-1} + \beta_2 e_{t-2} + ... + \beta_q e_{t-q}$$

$$X_t - \alpha_1 B X_t - \alpha_2 B^2 X_t - ..... - \alpha_p B^p X_t = e_t + \beta_1 B e_t + \beta_2 B^2 e_t + ... + \beta_q B^q e_t$$

$$(1 - \alpha_1 B - \alpha_2 B^2 - ..... - \alpha_p B^p)X_t = (1 + \beta_1 B + \beta_2 B^2 + ... + \beta_q B^q)e_t$$

$$\phi(B)X_t = \theta(B)e_t \quad (1.21)$$

Where $\phi(B)$ and $\theta(B)$ are polynomials of orders p and q respectively such that:

$\phi(B) = 1 - \alpha_1 B - \alpha_2 B^2 - ..... - \alpha_p B^p$

and

$\theta(B) = 1 + \beta_1 B + \beta_2 B^2 + ... + \beta_q B^q$

1.5.7 Autoregressive Integrated Moving Average -ARIMA(p,d,q) Process

The ARIMA model is a combination of two univariate time series models which are Autoregressive (AR) model and Moving Average (MA) model.

We say that $\{X_t\}$ is an ARIMA process of order (p,d,q) written as $X_t \sim ARIMA(p, d, q)$ if the d^{th} difference of $\{X_t\}$ is a stationary and invertible ARMA process of order (p,q). By using the backward shift operator the ARIMA process is given as:

$$\phi(B)(1 - B)^d X_t = \theta(B)e_t \quad (1.22)$$

Where $e_t \sim WN(0, \sigma^2)$ and $\phi(B)$ and $\theta(B)$ are polynomials of degrees p and q respectively with all the roots of the $\phi(B) = 0$ and $\theta(B) = 0$ lying outside the unit circle

Kleiber and Zeileis [16] assert that the ARIMA model is applied in the case where the series is non-stationary and an initial differencing step (corresponding to the "integrated" part of the model) can make ARMA model applicable to an integrated stationary process. consequently, the ARIMA model with its order is presented as ARIMA (p,d,q) model where p, d, and q are integers greater than or equal to zero and refer to the order of the autoregressive, integrated, and moving average parts of the model respectively. The first parameter p refers to the number of autoregressive lags (not counting the unit roots), the

second parameter d refers to the order of integration that makes the data stationary, and the third parameter q gives the number of moving average lags.

1.5.8 Seasonal Autoregressive Integrated Moving Average

A time series $\{X_t\}$ is said to be seasonal if there exists a tendency for the series to exhibit a periodic behaviour after certain time interval. The usual ARIMA models cannot really cope with seasonal behaviour, it only models time series with trends. Seasonal ARIMA models are formed by including an additional seasonal terms in the ARIMA models and are defined by seven parameters p,d and q which are the order of non seasonal AR, differencing and MA respectively; P,D and Q which are the order of seasonal AR, Differencing and MA respectively and S which represents seasonal order.

Therefore a $SARIMA(p,d,q)(P,D,Q)_S$ model can be written as:

$$\phi(B)\Phi(B^S)(1-B)^d(1-B^S)^D X_t = \theta(B)\Theta(B^S)e_t \tag{1.23}$$

where;

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

$\Phi(B^S) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - ..... - \Phi_p B^{pS}$

$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - ..... - \theta_q B^q$

$\Theta(B^S) = 1 - \Theta_1 B^S - \Theta_2 B^{2S} - .... - \Theta_q B^{qS}$

e_t is a purely random process

1.5.9 Time series forecasting

It is a planning tool which helps decision makers to foresee the future uncertainty based on the behaviour of past and current observations. Forecasting as described by Box and Jenkins [3], is the process of predicting some unknown quantities.

Let X_1,X_2,.....X_n be observed time series. Forecasting involves determining the future value say X_{n+k} made at time n for k steps ahead, k = 1, 2, ... and is called the lead time.

The forecast of X_{n+k} made at time n for k steps ahead denoted by $\widehat{X}_(n,k)$ is the value for which the mean squared error (MSE)of the predictor $\widehat{X}_(n,k)$ is minimum i.e.

$$MSE[X_{n,k}] = E[X_{n+k} - \widehat{X}_(n,k)]^2 \tag{1.24}$$

should be minimum

Chapter 2

Review of Related Literature

2.1 Introduction

This section reviews literature related to the application of seasonal ARIMA in modelling time series data.

2.2 Review of Related Literature

As observed by Janine et al.(2004) Univariate time series modelling techniques are becoming an increasingly popular method of analyzing time series data. These techniques include models such as Autoregressive (AR), Moving Average (MA), and Autoregressive Integrated Moving Average (ARIMA). Further, they assert that ARIMA models are theoretically justified and can be very robust with respect to alternative (multivariate) modeling approaches with respectable forecast performance relative to the theoretically based specifications. Box and Jenkins [3] provided an approach of ARIMA model building which includes model identification, estimation, diagnostic checking, and forecasting a univariate time series. According to Caldwell [5], the Box-Jenkins methodology is particularly suited for development of models of processes exhibiting strong seasonal behaviour.

Maamar [17] in his study 'ANN versus SARIMA models in forecasting residential water consumption in Tunisia' used quarterly time series household water consumption data and did a comparative analysis between the traditional BoxJenkins method and an

artificial neural networks approach. In particular, they attempted to test the effectiveness of data pre-processing, such as detrending and deseasonalization, on the accuracy of neural networks forecasting. Results indicate that the traditional BoxJenkins method outpeforms neural networks estimated on raw, detrended, or deseasonalized data in terms of forecasting accuracy. They established that forecasts provided by the neural network model estimated on combined detrended and deseasonalized data are significantly more accurate and much closer to the actual data. The model was therefore selected to forecast future household water consumption in Tunisia with Projection results suggesting that by 2025, water demand for residential end-use will represent around 18 percent of the total water demand of the country.

Habib et al. [10] carried out a study on 'Estimation of Water Demand in Iran Based on SARIMA Models' in which they employed the seasonal autoregressive integrated moving average (SARIMA) model. They fitted this model to monthly residential water consumption in Iran from May 2001 to March 2010. They established that a three-parameter log-logistic distribution fits the model residuals adequately. They forecast values for 1 year ahead using the fitted SARIMA model. They established the best fitting models to be $SARIMA(1,1,0)(1,1,0)_6$, $SARIMA(0,1,1)(1,1,0)_6$ and $SARIMA(0,1,1)(0,1,1)_6$ with residuals described by a $3-$parameter $\log-$logistic distribution.

Bithas and Stoforos [1] applied single equation econometric analysis through a regression analysis to estimate domestic water use. The policy-relevant variables, mainly income and water prices, were systematically considered and their effects on water demand appraised. The study established that a drastic increase in water demand induced by increasing income will occur, while the economic instruments have little potential to influence water use. The factors real GDP, Real water price and trend had a significant influence on residential water demand with $R^2 = 0.91$. Trend which was used as a proxy for weather variation was found to be more important than price. They applied the root mean square percent error(RMSPE) and Mean Percent Error (MPE) to asses the model forecast accuracy. The values 0.01 and -0.01 obtained for RMSPE and MPE respectively indicated that the model developed tracked historical development in water demand fairly

well.

Smaoui et al.[29] used two approaches, namely Box−Jenkins approach and artificial neural networks approach (ANN) to model time series data of water consumption in Kuwait. The Box-Jenkins approach was used to predict unrecorded water consumption data from May 1990 to December 1991 due to the Iraqi invasion of Kuwait in August 1990. A supervised feed forward back propagation neural network was then designed, trained and tested to model and predict water consumption from January 1980 to December 1999. It is interesting to note that the lagged or delayed variables obtained from the Box-Jenkins approach and used in neural networks provide a better ANN model than the one obtained either blindly in black box mode as has been suggested or from traditional known methods. It was found that when the variables of the input layer in ANN is chosen based on the Box−Jenkins approach rather than on traditional methods, the average relative error for the training and testing data sets are reduced by 24 percent hence the assertion that the combination of Box−Jenkins approach and ANN is superior in predicting the water consumption than the ANN alone.

Martinez−Espineira [20] carried out a study in which residential water demand were estimated using co−integration and error-Correction methods based on monthly time series observations from Seville (Spain). Unit root tests revealed that water use series and series of other variables affecting use were non−stationary. However, a long-run co-integrating relationship was found in the water demand model. This made it possible to obtain a partial correction term and to estimate an error correction model. The price-elasticity of demand was estimated at around -0.1 in the short run and -0.5 in the long run.

Schleich and Hillebrand [28] econometrically analyzed the impact of several economic, environmental and social determinants for the per capita demand for water in about 600 water supply areas in Germany. Besides prices, income and household size, they considered the effects of population age, the share of wells, housing patterns, precipitation and temperature. They also explored why current per capita residential water consumption in

the new federal states was about 3 percent lower than in the old federal states. Since average cost pricing may cause an endogeneity problem, they applied instrumental-variable procedures in addition to single equation ordinary least squares, but found no evidence that prices were endogenous. Their estimation results suggested that the price elasticity of water demand in Germany was around 0.24. The income elasticity was positive, decreased with higher income levels and was at least three times higher in the new federal states than in the old federal states. Differences in prices and income levels explained about one third of the gap in residential water use between the two regions. Household size and the share of wells had a negative impact on per capita water demand, and water use increased with age. Finally, the findings provided some evidence that rainfall patterns rather than total rainfall affected water consumption, while temperature appears to have no impact at all. All outcomes were robust to a loglog and two types of semi−log specifications for the water demand function.

Xinming, Dale and Briscoe [37] applied Ordinary Least Squares to model water demand in Ukunda, Kenya. Water consumed per capita per day (lcd) was hypothesized to be a function of the independent variables: time, income, education, number of adult women in the households, number of water vendors and number of water kiosks. The OLS estimations showed that only the number of women in the households as a proportion of total household size had a significant effect on demand at 0.05 level of significance. All the other variables were found not to be significant even at 0.1 level of significance. The finding that the level of income had no effect on water demand negates a finding by [32] who applied Pearson Correlation Analysis to test for the strength of the relationship between monthly household income and water and obtained an r value of 0.992 indicating a strong positive correlation between household income and daily per capita water use.

Yaw [38] in his PHD thesis on household water security and water demand in the Volta basin of Ghana established through OLS procedure that Household size, price of improved water and the gender of the household head were significant determinants of water demand at 1 percent. Further the study showed that the larger the household size, the higher the consumption of improved water where a 10 percent increase in household

size led to approximately 4 percent rise in improved water demand. The demand for improved water was found to be price inelastic where the effect of a 10 percent increase in price decreased quantity demanded by 3.6 percent. Household income was established to have a positive relationship with improved water demand but with a weak effect. This result suggested that household income as a decision variable that influences the probability of using improved water was important but was unimportant in determining quantities consumed thereof.

Buckman and Mintah[4] applied Autoregressive Integrated Moving Average (ARIMA) to Model Ghanas monthly inflation from January 1985 to December 2011 and used the model to forecast twelve (12) months inflation for Ghana. Using the Box Jenkins (1976) framework, the autoregressive integrated moving average (ARIMA) was employed to fit the best ARIMA model. The seasonal ARIMA model, SARIMA (1, 1, 2) (1,0, 1) was chosen as the best fitting from the ARIMA family of models with least Akaike Information Criteria (AIC) of 1156.08 and Bayesian Information Criteria (BIC) of 1178.52. The plots of actual values and the forecasted values of inflation were very close implying that the selected model best fit the data and hence, appropriate for forecasting. The forecast error 3.4 also gave further evidence that the model selected had very strong predictive power.

Osabuohien [24]in his empirical study which aimed at modelling and forecasting time series quarterly data of Rainfall in Port−Harcourt, south Nigeria using the Box-Jenkins SARIMA Methodology established that the seasonal model$ARIMA(0,0,0)x(2,1,0)_4$ fitted to the series appropriately. A forecast from 2009 to 2013 was made and the forecast obtained on the basis of the fitted model was adequate.

Mahsin et al.[18] applied Box-Jenkins methodology in modelling rainfall in Dhaka Division in Bangladesh. They build a seasonal ARIMA model for monthly rainfall data taken for the period from 1981-2010 (June) with a total of 354 readings. The model $ARIMA(0,0,1)(0,1,1)_{12}$ was found adequate and was used to forecast the monthly rainfall for the upcoming two years to help decision makers to establish priorities in terms

of water demand management. The RMSE values on test data were comparatively less hence the prediction model was considered reliable. By comparing the fitted and actual values of rainfall data using the determined model the rainfall forecasts made for the years 2011 and 2012 were sufficiently accurate at 95 percent confidence interval.

2.3 Forecasting Time Series

Most forecasting problems involve the use of time series data. Montgomery et al [21] stated that forecasting problems are often classified as short-term, medium term, and long-term. Short-term forecasting problems involve predicting events only a few time periods (days, weeks, months) into the future. Medium-term forecasts extend from one to two years into the future, and long-term forecasting problems can extend beyond that by many years. Short-term and medium-term forecasts are used for operations management and development of projects while long-term forecasts can be used for strategic planning.

Normally, short-term and medium-term forecasts are based on identifying, modelling, and extrapolating the patterns found in historical data. These historical data usually exhibit inertia and do not change very drastically. Therefore, statistical methods are very useful for short-term and medium-term forecasting[21].

In summary, most of the studies carried out on the application of SARIMA have been focused on the modelling of rainfall, inflation and other variables. The fewer studies that have modelled residential water demand have used a combination of the Box−Jenkins approach and ANN. Others have used co-integration and error-Correction methods while others have applied SARIMA. The few studies carried out in the Kenyan context have applied Ordinary Least Squares to identify the determinants of water demand. Little or no studies have modelled residential water demand by applying SARIMA and therefore this study comes in handy to fill this void.

Chapter 3

Research Methodology

3.1 Introduction

This chapter presents research design, location of the study, data collection procedure, trend analysis, water demand modelling technique and water demand model building procedures. The Box−Jenkins approach will be applied to develop the residential water demand SARIMA model which will be applied to forecast water demand for the 12 months of 2014.

3.2 Research Design

The research employed a case study design. This research design was considered appropriate because the focus was on a functioning specific individual unit (KIWASCO) with set boundaries [30].

3.3 Study Area

The proposed research examined water demand analysis in Kisumu City which is located on the eastern shores of Lake Victoria at the tip of Winam Gulf and it is the third largest city in Kenya. The city covers a total area of 417 sq. km, of which 297 sq. km is land, and 120 sq. km is water mass. The city, which has been designated as a regional growth

node, is connected to nation and the region by four major roads. The major routes are Nairobi Road to the southwest of the town, which connects Kisumu to Nakuru, Nairobi and Mombasa. To the north is a connection to Kakamega while to the west is a connection to Busia. The Busia route provides an alternative road to Uganda via Kisumu. The forth road into Kisumu is a small connection to Kibos and Muhoroni to the east of the town. Kisumu is also connected to Nairobi and Mombasa by a major rail link. It has a strategic position in the East African Cooperation which is currently under consideration due to its accessibility to the regional countries; Uganda, Tanzania, Rwanda and Burundi.

The city is located in a place with intensified scarcity; frequent and lasting drought periods and rapid expanding urbanization. It experiences warm to hot and generally humid climate with monthly maximum and minimum temperatures varying from 28^0C to 31^0C and 16^0C to 18^0C, respectively. Higher and lower temperatures are experienced during October to March and April to August, respectively. The economic and demographic growth in the last decades has transformed the city into an important industrial and commercial center. The city has an international airport, several universities and an extensive public health network. The accessibility to water services is below the national average despite close proximity of World's second largest fresh water lake.

The current water production and distribution system is predominantly run by Kisumu Water and Sewerage Company (KIWASCO), which is the main water provider with the mandate to provide water within the region and is therefore focused as managers of the present and future water facilities. Kajulu water treatment works and Dunga water treatment works with River Kibos and Lake Victoria as the water sources respectively. The main distribution reservoirs are located at Tom Mboya estate within Kibuye Sub-location.

3.4 Data Collection Procedure

The monthly residential water demand data were obtained from the Kisumu Water and Sewerage Company (KIWASCO) records. KIWASCO, is the main water provider with the mandate to provide water within the region and is therefore considered as manager of the present and future water facilities. The data observed from January 2004 to December

2013 with 120 data entries was used to develop the SARIMA model. The monthly amount of water consumption during the year 2013, was used to validate the SARIMA model.

3.5 Residential water demand trend analysis for the period 2004−2013

3.5.1 Descriptive Statistics

As a prelude to the application of linear regression analysis on the residential water demand, descriptive analysis both numerical and graphical was first be carried out in order to explore the residential water demand situation in Kisumu City. The arithmetic mean was used for numerical summary measure, whereas scatter plots and trend lines of residential water against time were generated for graphical presentation of the data.

3.5.2 Testing of Trend

For comparative purposes of the results in this study, both parametric and non−parametric methods were done. Parametric tests are more powerful when the data are normally distributed than is the case when it is not [23]. When the data include outliers or are severely non-normally distributed, the use of parametric methods can give incorrect results hence invalid inference. Moreover, according to Racine [26], non−parametric methods relax the parametric assumption imposed on the data generating process and allow the data to determine a suitable functional form. Ordinary Least Squares (parametric test) and Kendall's Tau Test(non−parametric test) will be used to test trend.

Ordinary Least Squares test of trend

Before the linear regression procedure the normality test was carried out. This tested the likelihood that the water demand data set x_1.....x_n came from a Gaussian distribution. The Shapiro−Wilk test was used to achieve this. In statistics, the Shapiro−Wilk

test tests the null hypothesis that a sample $x_1.....x_n$ came from a normally distributed population. Therefore the hypotheses of this test were:

H_0: The monthly demand for water in Kisumu City is normally distributed

H_1: The monthly demand for water in Kisumu City is not normally distributed.

The test statistic is:

$$S = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.1}$$

where,

i) $x_{(i)}$ with parentheses enclosing the subscript index is the i^{th} order statistic, i.e.,the i^{th} smallest number in the sample

ii) $\bar{x} = \frac{x_1+.....+x_n}{n}$

iii) The constants a_i are given by $(a_1, a_2, ...a_n) = \frac{m^T V^{-1}}{\sqrt{m^T V^{-1} V^{-1} m}}$

where $m = (m_1....m_n)^T$ are the expected values of order statistics of independent and identically distributed random variables sampled from the standard normal distribution and V is the covariance matrix of those order statistics.

The Ordinary Least Squares equation relating the amount of water demanded with time was:

$$y = \gamma_0 + \gamma_1 t + \varepsilon \tag{3.2}$$

Where, y is the dependent or response variable representing the amount of water demanded monthly, t is the covariate or explanatory variable and ε is the unobserved error or disturbance. The goal will be to estimate the regression parameters the intercept i.e. γ_0 and the slope γ_1. A familiar assumption in linear regression is that the error has a mean of zero and that each explanatory variable is uncorrelated with the error term [33]. In the structure of the model in Equation 3.5.2 this assumption is equivalent to $E(\varepsilon) = 0$,

$E(t, \varepsilon) = 0$.

In the perspective of the present analysis, γ_1 was interpreted as representing the average rate of change of water demand throughout every one month time period. Significant γ_1 ($p < 0.05$) is an indicator of trend in the amount of water demanded. Otherwise, insignificant γ_1 signify absence of trend in the amount of water demanded over time. Also,a negative sign of γ_1 indicates a decreasing trend while a positive value imply an increasing trend with time.

The least squares estimate of $\widehat{\gamma_0}$ and $\widehat{\gamma_1}$ are obtained using the following equations:

$$\begin{aligned} \hat{\gamma}_1 &= \frac{\sum_{i=1}^{n} y_i t_i - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} t_i}{n}}{\sum_{i=1}^{n} t_i^2 - \frac{(\sum_{i=1}^{n} t_i)^2}{n}} \\ \textit{and} & \\ \hat{\gamma}_0 &= \bar{y} - \widehat{\gamma_1}\bar{t} \end{aligned} \tag{3.3}$$

Where, $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ which is the mean of the observed residential water demand
$\bar{t} = \frac{\sum_{i=1}^{n} t_i}{n}$ which is the mean of the predictor variable.

The hypothesis to be tested in this case were:

H_0: $\gamma_1 = 0$ i.e. the slope is equal to zero

H_a: $\gamma_1 \neq 0$ i.e. the slope is not equal to zero.

The test statistic used was the F-Test based on the Analysis of Variance (ANOVA) which is summarized in the table below:

Table 3.1: ANOVA Table for the test of trend

Source of Variation	df	Sum of Squares	Mean Squares	F-Value
Regression	1	$\frac{(S_{ty})^2}{S_{tt}}$	$\frac{(S_{ty})^2}{S_{tt}}$	$\frac{MS_{Regression}}{MS_{Error}}$
Error	n-2	$S_{yy} - \frac{(S_{ty})^2}{S_{tt}}$	$\frac{S_{yy} - \frac{(S_{ty})^2}{S_{tt}}}{n-2}$	
Total	n-1	S_{yy}		

where;

$SS_{tot} = S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\sum_{i=1}^{n}(y_i)^2}{n}$

$SS_{Regr} = \frac{(S_{ty})^2}{S_{tt}}$

$S_{ty} = \sum_{i=1}^{n} y_i t_i - \frac{\langle\sum_{i=1}^{n} y_i\rangle\langle\sum_{i=1}^{n} t_i\rangle}{n}$

$S_{tt} = \sum_{i=1}^{n} t_i^2 - \frac{\langle\sum_{i=1}^{n} t_i\rangle^2}{n}$

$SS_{Error} = S_{yy} - \frac{(S_{ty})^2}{S_{tt}}$

H_0 will be rejected if $F_{Calc} > F_{(1,n-1)(\alpha=0.05)}$

As asserted by Gupta [9], the line obtained by OLS is the line of best fit because it is the line from where the sum of positive and negative deviations is zero i.e. $\sum(y - y_c) = 0$ and the sum of the squares of the deviations i.e.$\sum(y - y_c)^2$ is least

Kendall's Tau test of trend

According to Onoz and Bayazit [23], the Kendall's τ statistic is one of the non−parametric trend tests that have been frequently used and is considered an excellent reference for numerous other trend test techniques. Kendal's τ test first ranks all observations by date order, then the difference between each consecutive value is calculated and the sum of

the signs of these differences is calculated as the Kendall sum, S statistic given as in the Equation below:

$$S = \sum_{i=k}^{n-1} \sum_{i=k+1}^{n} Sgn(X_i - X_k) \tag{3.4}$$

Where;

$$Sgn(X_i - X_k) = \begin{cases} 1 & \text{if } X_i - X_k > 0 \\ 0 & \text{if } 0 < h < q \\ -1 & \text{if } X_i - X_k < 0 \end{cases}$$

The expected value and variance S are:

$E(S) = 0$

and

$$Var(S) = \frac{[n(n-1)(2n+5) - \sum_t (t(t-1)(2t+5)]}{18} \tag{3.5}$$

t indicates the extent of any given time and $\sum_t$ denotes the sum across all ties in the water demand data.

For $n > 0$, the standard normal variate is given by:

$$Z = \begin{cases} \frac{S-1}{\sqrt{Var(S)}} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ \frac{S+1}{\sqrt{Var(S)}} & \text{if } S < 0 \end{cases}$$

Under the Kendall's τ test, a positive value of S in Equation (3.4) indicates an increasing trend whereas a negative value indicates a decreasing trend [13]. The null and alternative hypotheses under S are:

H_0: There is no trend in the monthly demand for water.

H_a: There is a trend in the monthly demand for water.

The decision is to reject the null hypothesis if the p−value of the test is less than the level of significance. Kendall's τ test of significance will be two-sided and considered significant at the 0.05 level. All the test statistics and figures were generated using EViews 8 and Minitab Version 17 softwares.

3.5.3 Testing for Stationarity and Autocorrelation

Before the search for the best model for the data, the first condition was to check whether the series is stationary or not. The SARIMA model is appropriate for stationary time series data (i.e. the mean, variance, and autocorrelation are constant through time). If a time series is stationary then the mean of any major subset of the series does not differ significantly from the mean of any other major subset of the series. Also if a data series is stationary then the variance of any major subset of the series will differ from the variance of any other major subset only by chance [25]. The stationarity condition ensures that the autoregressive parameters in the estimated model are stable within a certain range as well as the moving average parameters in the model are invertible.

To check for stationarity, we usually test for the existence or non existence of unit root. Unit root test is performed to determine whether a stochastic or a deterministic trend is present in the series. If the roots of the characteristic equation lie outside the unit circle, then the series is considered stationary [7]. Kwiatkowski−Phillips−Schmidt−Shin(KPSS) and the Augmented Dickey Fuller (ADF)tests were used to test for stationarity in the data. The null hypothesis for the ADF test is that the water demand series have unit roots or the series is non-stationary. The null hypothesis is rejected if the test statistic is larger in the absolute term than the critical value. The Augmented Dickey-Fuller (ADF) test was performed to determine whether data differencing was needed[7].

Also using KPSS test, the study tested the null hypothesis that the original series is stationary at the non seasonal level

The KPSS hypothesis may be stated as:

$H_0 : \sigma^2 = 0$,(Stationary)

$H_a : \sigma^2 \neq 0$,(non−Stationary)

The KPSS test statistic is given by:

$$KPSS = \frac{(T^{-2}\sum_{t=1}^{T}\hat{S}_t^2)}{\hat{\lambda}^2} \tag{3.6}$$

The autocorrelations were determined by computing the ACF. Given that $\{X_t\}$ is a stationary time series, with constant expectation and time independent covariance. The ACF for the series is defined as:

$$\rho_h = \frac{Cov(X_t, X_{t+h})}{\sqrt{Var(X_t)Var(X_{t+h})}} = \frac{\gamma(h)}{\gamma(0)} \tag{3.7}$$

for $h > 0$. The value h denotes the lag.

Plots of ACF as a function of h were done, and this helped to determined if the autocorrelation decreases as the lag gets larger or of if there is any particular lag for which the autocorrelation is large

3.6 Water Demand Modelling Technique

Let X_1, X_2.....X_n represent a sequence of seasonal observations representing monthly residential water demand. To eliminate non−stationarity within each season, one can employ the seasonal differencing operator $(1 - B^S)$ resulting to:

$$\nabla_S X_t = (1 - B^S)X_t = X_t - X_{t-S} \tag{3.8}$$

where t=S+1, S+2.....

S is the number of seasons per year

and B^S is the backward shift operator

If we apply the seasonal differencing operator in equation (3.8) D times we produce a

series given by:

$$\nabla_S^D X_t = (1 - B^S)^D X_t \tag{3.9}$$

To model correlation between same months observations in the differenced series, one may wish to introduce appropriate model parameters. To accomplish this, one can use a model of the form:

$$\Phi(B^S)\nabla_S^D X_t = \Theta(B^S)e_t \tag{3.10}$$

where $\Phi(B^S)$ and $\Theta(B^S)$ are the seasonal autoregressive(AR)and seasonal moving average(MA) operators, respectively and e_t is a residual series which may contain non −seasonal correlation. Both the seasonal AR and MA operators are defined in order to describe the relationship within the same season.
In particular the seasonal AR operator is defined as:

$$\Phi_P(B^S) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - ..... - \Phi_P B^{PS} \tag{3.11}$$

Φ_i is the i^{th} AR parameter and P is the order of the AR operator.

Since the power of each differencing operator is always an integer multiplied by S, only observations within each season are related to one another when using this operator. Same months observations are connected together by $\Phi_P(B^S)$.
To describe the relationship of the residuals e_t within a given season, the seasonal MA operator is defined using:

$$\Theta_Q(B^S) = 1 - \Theta_1 B^S - \Theta_2 B^{2S} - .... - \Theta_q B^{qS} \tag{3.12}$$

where, Θ_i is the i^{th} MA parameter and Q is the order of the MA operator.
Since the power of each differencing operator is always an integer multiplied by S, the residuals in the same season are linked to one another when using the operator $\Theta_Q(B^S)$. The residuals e_t may contain non −seasonal non−stationarity which can be removed using the non−seasonal differencing operator defined by:

$$\nabla^d e_t = (1 - B)^d e_t \tag{3.13}$$

Where d is the order on the non −seasonal differencing operator which is selected just large enough to remove all of the non −seasonal stationarity.

The sequence produced by using (3.13) is theoretically a stationary non −seasonal series. the non −seasonal correlation can then be captured by writing the ARMA model as:

$$\phi(B)\nabla^d e_t = \theta(B)e_t \tag{3.14}$$

Where $\phi(B)$ is the non −seasonal AR operator of order p defined as :

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - .....\phi_p B^p \tag{3.15}$$

and $\theta(B)$ is the non−seasonal MA operator of order q written as:

$$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - ..... - \theta_q B^q \tag{3.16}$$

The $e_t's$ are innovations that are IID with a mean of Zero and variance σ^2
To define the overall seasonal autoregressive integrated moving average model we combine (3.14) and (3.10). This is accomplished by solving for $e_t's$ in (3.14) and substituting this result into (3.10) to obtain the SARIMA model.

$$\phi_p(B)\Phi_P(B^S)\nabla^d\nabla_S^D X_t = \theta_q(B)\Theta_Q(B^S)e_t$$

$$OR$$

$$\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D X_t = \theta_q(B)\Theta_Q(B^S)e_t \tag{3.17}$$

According to Halim and Bisono [11], an economical notation for summarising the structure of the SARIMA model is:

$$SARIMA(p,d,q)(P,D,Q)_S \tag{3.18}$$

Where p and P are the orders of autoregressive operator of non-seasonal and seasonal components respectively; d and D are the differences of non-seasonal and seasonal components respectively and q and Q are the orders of moving average operator of non-seasonal and seasonal components respectively. S is the seasonal length.

3.7 Water Demand Model Building

Water demand modelling was based on the Box−Jenkins approach. Based on Caldwell[5], the Box−Jenkins approach was preferred because of its capability to capture the appropriate trend by examining historical pattern; it is helpful in extracting a great deal of information from the time series using a minimum number of parameters and has the capability of handling stationary and non−stationary time series in non-seasonal and seasonal elements.

There are four main stages in building an ARIMA model based on Box-Jenkins procedure[3], i.e., (1) model identification, (2) model estimation, (3) model checking and (4) model forecasting. These stages of building an ARIMA model are described in the figure 3.1 below.

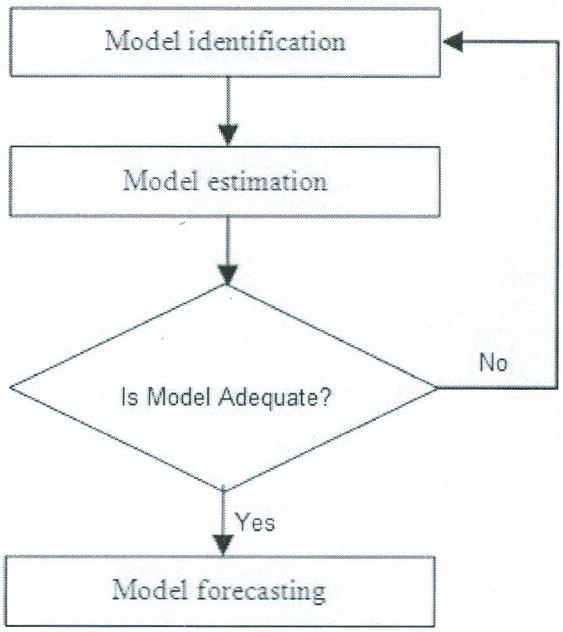


Figure 3.1: Box-Jenkins Modelling Procedure

3.7.1 Model Identification

When the series is stationary, the order of the model which is the AR, MA, SAR and SMA terms can be determined. Where AR=p and MA=q represent the non-seasonal autoregressive and moving average parts respectively and SAR=P and SMA=Q represent the seasonal autoregressive and moving average parts respectively as described earlier. To determine these orders, the study made use of the sample autocorrelation function (ACF) and partial autocorrelation function (PACF) of the stationary series. The ACF gives information about the internal correlation between observations in a time series at different distances apart, usually expressed as a function of the time lag between observations. These two plots suggest the model we should build. Checking the ACF and PACF plots, we should both look at the seasonal and non-seasonal lags. Usually the ACF and the PACF has spikes at lag k and cuts off after lag k at the non-seasonal level. Also the ACF and the PACF has spikes at lag ks and cuts off after lag ks at the seasonal level. The number of significant spikes suggests the order of the model.

3.7.2 Parameter Estimation

Maximum Likelihood Method under the normal distribution was applied to estimate the model's parameters . This will involve choosing values for the parameters that maximizes the likelihood of the data occurring. Given a sample $x_1, x_2....x_n$ of n, IID observations, which comes from a distribution f(x) with unknown parameter θ , then; the joint density function is

$$f(x_1, x_2....x_n, \theta) = f(x_1, \theta)(*f(x_2, \theta) * ..... * f(x_n, \theta)$$

By considering the observed values $x_1, x_2....x_n$ to be fixed parameters of this function, whereas θ will be the function's variable and allowed to vary freely. And this function is called likelihood

$$L(\theta, x_1, x_2....x_n) = f(x_1, x_2....x_n, \theta)$$

$$=\prod_{i=1}^{n} f(x_i, \theta)$$

In practice, it is often more convenient to work with the logarithm of the likelihood function and called the log-likelihood:

$Ln(\theta, x_1, x_2....x_n) = \sum_{i=1}^{n} Lnf(x_i, \theta)$

Parameter Estimation for SARIMA (p,d,q)(P,D,Q)S model

Different SARIMA models will be applied to find the best fitting model. The most appropriate model will be selected by using the Bayesian information criterion (BIC) and Akaike information criterion (AIC) values. The best model will be determined from the minimum BIC and AIC. Thus, the minimization of AIC or BIC is more satisfactory for choosing the best model from candidate models having different numbers of parameters.

$$AIC = nln(\frac{S}{n}) + 2p \tag{3.19}$$

and

$$BIC = nln(\frac{S}{n}) + p + pln(n) \tag{3.20}$$

where n is the number of effective observations used in fitting the model; p is the number of parameters fitted in the model and S is the sum of squared residuals up to time T. In the two equations the first term $nln(\frac{S}{n})$ is a measure of "lack of fit" and the remainder is a penalty for increasing the number of model parameters.

3.7.3 Model Adequacy Checking

After estimating the parameters of ARIMA model, the next step in the Box-Jenkins approach is to check the adequacy of that model which is usually called model diagnostics and involves performing Goodness−of−fit tests based on the standardized residuals. Ideally, a model should extract all systematic information from the data. The residuals should be small. The diagnostic check is used to determine the adequacy of the chosen model. One assumption of the ARIMA model is that, the residuals of the model should be white noise. A series $\{X_t\}$ is said to be white noise if $\{X_t\}$ is a sequence of independent and identically distributed random variable with finite mean and variance. In addition

if $\{X_t\}$ is normally distributed with mean zero and variance , σ^2 then the series is called Gaussian White Noise. For a white noise series the, all the ACF are zero.

The normal plots, ACF and the PACF plots of residuals were done to check for model adequacy. The normal probability plot should be a straight line while the time plot should exhibit random variation. For ACF all the correlation should be within the test bounds which indicates stationarity in the data.In practice if the residuals of the model is white noise, then the ACF of the residuals are approximately zero. The sample autocorrelation function (ACF) and partial autocorrelation function (PACF) are useful qualitative tools to assess the presence of autocorrelation at individual lags. The Shapiro normality test will also be used to check for homoscedasticity and normality among the residuals. Montgomery [22] determined that, if the model is adequate, the residuals should be structure-less, that is, they should contain no obvious patterns. However, a very common defect that often shows up on the normal probability plots is one residual being much larger than the others, and this can seriously distort the analysis of variance.

Following[2], Ljung-Box test was employed to check for adequacy of the fitted model. The Ljung-Box test is a type of statistical test which tests whether any group of autocorrelations of a time series is different from zero. It performs a lack-of-fit hypothesis test for model specification, which is based on the Q-statistic.

$$Q = N(N+2)\sum_{h=1}^{m}\frac{\rho_h^2}{N-h} \tag{3.21}$$

where N is the length of the observed time series, h= number of auto-correlation lags included in the statistic, m is the number of lags being tested and ρ_h^2 is the square of the sample autocorrelation coefficient at lag h.

Under the null hypothesis of no serial correlation, the Q−statistic is asymptotically Chi-Square distributed and the null hypothesis that all ρ_h are zero is rejected if the value of the computed Q is larger than the critical Q−statistic from the chi-square distribution at the given level of significance. Alternatively, if the p−value is smaller than the conventional significance level, the null hypothesis that there are no autocorrelation will be rejected.

3.7.4 Validation of SARIMA Model

The forecasted residential water demand by using SARIMA model were compared with the observed Residential water demand for the 12 month data of 2013. In evaluating the sample forecasting capability of the SARIMA model, the root mean square error (RMSE), and the mean absolute percentage error (MAPE) were used. The model MAPE and RMSE values were compared with the MAPE and RMSE values of the OLS equation.

Root mean square error (RMSE)

The Root Mean Square Error (RMSE) is a frequently used measure of the difference between values predicted by a model and the values actually observed from the environment that is being modelled. These individual differences are also called residuals, and the RMSE serves to aggregate them into a single measure of predictive power.

The RMSE of a model prediction with respect to the estimated variable X_{model} is defined as the square root of the mean squared error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i} - X_{model,i})^2}{n}} \tag{3.22}$$

Where; $X_{obs,i}$ are the observed values at time i

$X_{model,i}$ are the observed values at time i

RMSE has the advantage of giving more weight to large deviations, thereby punishing uneven predictors.

Mean Absolute Percentage Error (MAPE)

The mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD), measures the accuracy of a method for constructing fitted time series values in statistics.

$$MAPE = \frac{100}{N} X \sum_{i=1}^{N} \left| \frac{(A_t - F_t)}{A_t} \right| \tag{3.23}$$

Where;

A_t is the actual observed time series

F_t is the estimated or forecast time series

N is the number of non-missing data points

Model with a minimum of these statistics i.e. RMSE and MAPE was to be considered to be the best for forecasting.

3.8 Forecasting Residential Water Demand for 2014

After a model has passed the entire diagnostic test, it becomes adequate for forecasting. Forecasting is the process of making statements about events whose actual outcomes have not yet been observed. It is an important application of time series. The statistical software Minitab Version 17 was used to forecast the residential water demand for the 12 months of 2014.

3.9 Data Presentation

The results of the analysis were presented in tables and graphs such as the trend plots, ACF and PACF plots, probability plots and Histogram with normal curve plots

Chapter 4

Results and Discussion

4.1 Introduction

This chapter presents the results of the analysis of data.

4.2 Statistical Parameters of the Water Use Data

The result of the descriptive statistics as shown in table 4.1 below show that the minimum water demand was $64270M^3$ experienced in February 2008 whereas the maximum value was $147469M^3$ and was experienced in July 2013. The mean residential water demand for the 120 months studied was $93260m^3$ with the other values deviating from the mean by $23160m^3$. Standard Error Mean was 2114. The skewness index was 0.85 and this indicated that the data exhibited positive skewness with more observations lying below the mean than above the mean.

Table 4.1: Statistical parameters of water demand data

YEAR	Min.	Max.	Mean	SE Mean	StDev	Skewness
2004	74923	118155	95851	4401	15246	0.50
2005	66341	82940	73051	1534	5315	0.60
2006	66385	86908	76564	1921	6653	-0.36
2007	70605	86204	77839	1272	4406	0.20
2008	64270	80824	74022	1210	4190	-1.01
2009	71222	95761	80175	1964	6805	0.85
2010	74913	91334	84481	1488	5155	-0.56
2011	93055	121340	105941	2363	8187	0.25
2012	120852	137554	129939	1496	5183	-0.05
2013	118962	147469	134747	2617	9067	-0.27
Overall	64270	147469	93260	2114	23160	0.85

4.3 Analysis of Trend in the residential Water Demand Data for Kisumu City

The first objective of the study sought to analyse the trend of residential water demand in Kisumu City using monthly water demand data for the years 2003 to 2013. To achieve this objective, OLS was applied to develop a linear trend equation whose fit was compared to quadratic trend fit. The F-test based on ANOVA was used to test for the significance of γ_1. Also presence of seasonal trend was tested using Ratio-to-Trend method and Kendall's tau test for seasonal trend.

4.3.1 Test for Normality of the Residential Water Demand Data

Before the OLS procedure was applied, normality test based on the Shapiro–Wilk test was done. The Hypotheses tested were:

H_0: The residential monthly demand for water in Kisumu City is normally distributed

H_a: The residential monthly demand for water in Kisumu City is not normally distributed.

The results of the test are shown in table 4.2 below:

Table 4.2: Shapiro-Wilk test for Normality

W	0.929
p-value	$p < 0.010$

The results of the analysis show that for the Shapiro−Wilk test value of 0.929 the p-value is less than 0.05 hence we reject the null hypothesis. Therefore, it is concluded that the monthly residential water demand for Kisumu City do not come from a normal population. This is further demonstrated in figure 4.1 below which shows that the data does not fit well to a normal distribution but it is rather positively skewed as also indicated by the positive skewness index of 0.85 in table 4.1 above.

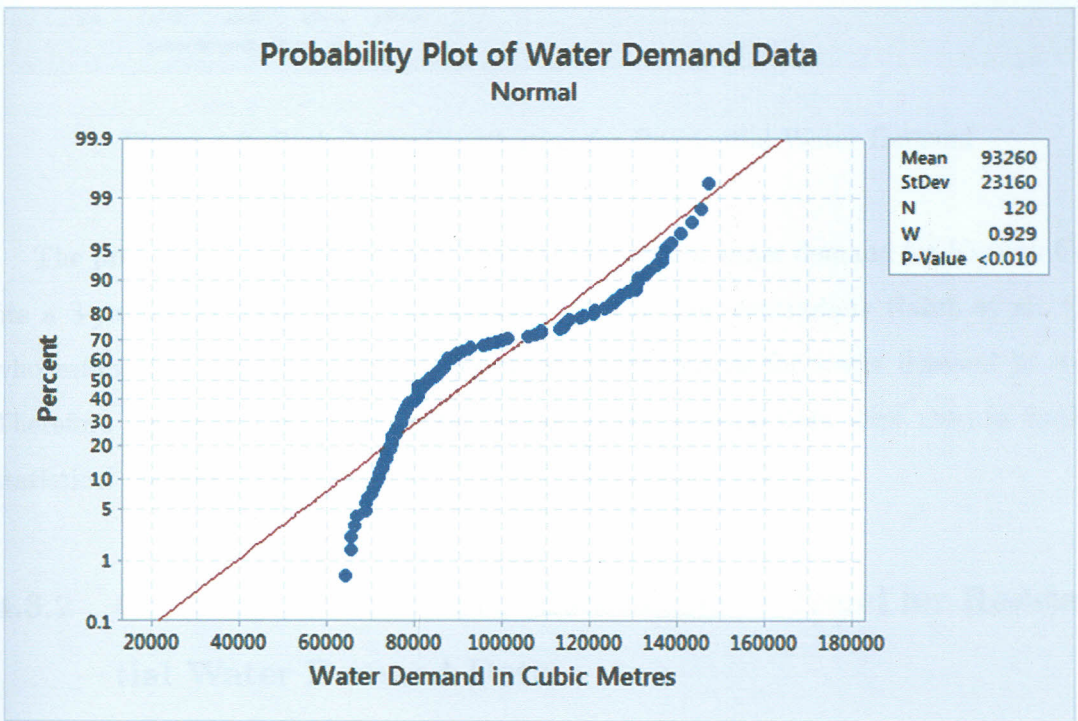


Figure 4.1: Normal Probability Distribution plot

Since the distribution of the residential water demand is not normal, the Series cannot be used for further statistical inferences hence the need to identify the best fitting distri-

bution. The plots for the various probability distributions of residential water demand are illustrated in the figure4.2 below

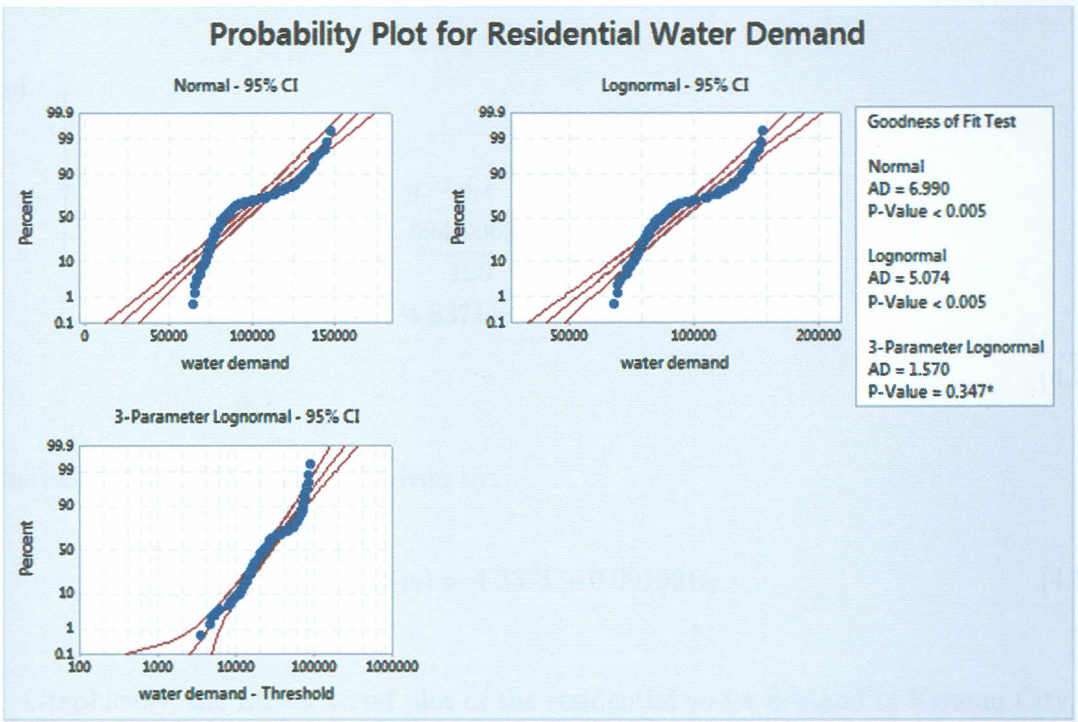


Figure 4.2: Probability Distribution for Residential Water Demand

The probability plots above show that the residential water demand for Kisumu City fits a 3-parameter log-normal distribution. This finding contradicts Habib et al. [10] who established a three-parameter log-logistic distribution for water demand in Iran. Therefore logarithmically transformed values of the original data were used in further statistical analysis.

4.3.2 Ordinary Least Squares Estimation of Trend for Residential Water Demand Data

The OLS equation is: $ln(y_i) = \gamma_0 + \gamma_1 t + \varepsilon$

The least squares estimates of $\hat{\gamma_0}$ and $\hat{\gamma_1}$ are:

$$\hat{\gamma_1} = \frac{\sum_{i=1}^{n} ln(y_i)t_i - \frac{\sum_{i=1}^{n} ln(y_i \sum_{i=1}^{n} t_i}{n}}{\sum_{i=1}^{n} t_i^2 - \frac{(\sum_{i=1}^{n} t_i)^2}{n}}$$

$$
\begin{aligned}
&= \frac{36278.46 - \frac{594.90635*7260}{120}}{583220 - \frac{(7260)^2}{120}} \\
&= 0.001991
\end{aligned} \tag{4.1}
$$

and

$$
\begin{aligned}
\hat{\gamma_0} &= \bar{y} - \hat{\gamma_1}\bar{t} \\
&= \frac{594.90635}{120} - 0.001991 * \frac{7260}{120} \\
&= 4.8371
\end{aligned} \tag{4.2}
$$

Therefore, the OLS equation is given by:

$$ln(\hat{y_i}) = 4.8371 + 0.001991t_i \tag{4.3}$$

Graphically, the Linear trend plot of the residential water demand in Kisumu City is illustrated in figure(4.3) below:

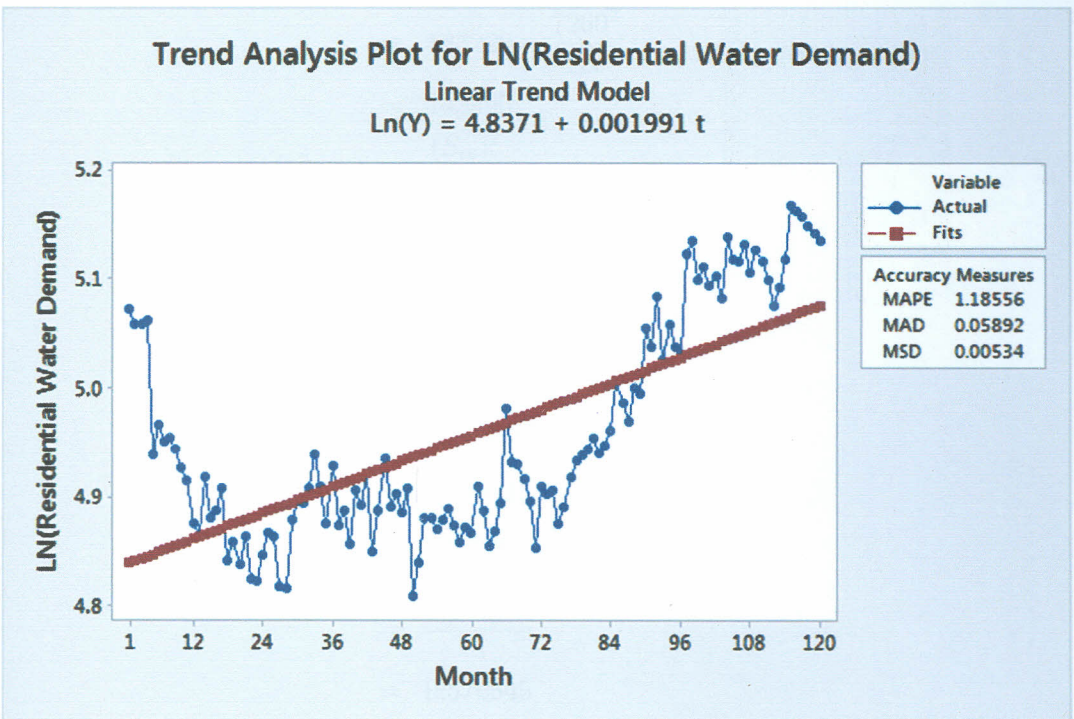


Figure 4.3: Linear Trend Analysis Plot

To test the significance of $\hat{\gamma}_1$, the F−test based on ANOVA was used. The hypotheses tested were:

H_0: $\gamma_1 = 0$ i.e. the slope is equal to zero

H_1: $\gamma_1 \neq 0$ i.e. the slope is not equal to zero.

Level of Signicance : $\alpha = 0.05$

Computations:

$$
\begin{aligned}
SS_{tot} = S_{yy} &= \sum_{i=1}^{n} lny_i^2 - \frac{(\sum_{i=1}^{n} lny_i)^2}{n} \\
&= 2950.491 - \frac{594.90635^2}{120} \\
&= 1.210963 \\
S_{ty} &= \sum_{i=1}^{n} lny_i t_i - \frac{\langle \sum_{i=1}^{n} y_i \rangle \langle \sum_{i=1}^{n} t_i \rangle}{n} \\
&= 36278.46 - \frac{594.90635 * 7260}{120} \\
&= 286.6231 \\
S_{tt} &= \sum_{i=1}^{n} t_i^2 - \frac{\langle \sum_{i=1}^{n} t_i \rangle^2}{n} \\
&= 583220 - \frac{7260^2}{120} \\
&= 143990 \\
SS_{Regr} &= \frac{(S_{ty})^2}{S_{tt}} \\
&= \frac{286.6231^2}{143990} \\
&= 0.570545 \\
SS_{Error} &= S_{yy} - \frac{(S_{ty})^2}{S_{tt}} \\
&= 1.210963 - 0.570545 \\
&= 0.640418 \\
MS_{Regr} &= \frac{SS_{Regr}}{df} \\
&= \frac{0.570545}{1} \\
&= 0.570545 \\
MS_{Error} &= \frac{SS_{Error}}{df}
\end{aligned}
$$

$$\begin{aligned} &= \frac{0.640418}{118} \\ &= 0.005427 \\ F-statistic &= \frac{MS_{Regr}}{MS_{Error}} \\ &= \frac{0.570545}{0.005427} \\ &= 105.13 \end{aligned} \tag{4.4}$$

The computations are illustrated in an ANOVA table as shown in table 4.3 below

Table 4.3: ANOVA Table for the test of trend

Source	df	SS	MS	F
Regression	1	0.570545	0.570545	105.13
Error	118	0.640418	0.005427	
Total	119	1.210963		

$F_{critical} = F_{(1,119)(\alpha=0.05/2)} = 5.15$

Since the $F_{Calc}(105.13) > F_{(1,119)(\alpha=0.05/2)}(5.15)$, the null hypothesis is rejected and the alternative hypothesis that the slope is not equal to zero is accepted at 0.05 level of significance. Based on the OLS procedure, it is therefore concluded that there is a significant positive(increasing) trend in residential water demand in Kisumu City as indicated by the positive $\gamma_1 = 0.001991$

The standard error of the estimate (Se) of $\hat{Y}$ is given by:

$$\begin{aligned} Se &= \sqrt{MSE} \\ &= \sqrt{0.005427} \\ &= 0.07367 \end{aligned} \tag{4.5}$$

Since the antilog of 0.07367 which is 1.185 means that the data is normally distributed with an interval of one standard deviation above the mean and one standard deviation below the mean encompassing 68.8 percent of all the observations.

4.3.3 Kendall's Tau Test

Under the Kendall's tau test for trend the hypotheses to be tested were;

H_0: There is no trend in the monthly demand for water.

H_1: There is a trend in the monthly demand for water.

The test results were as shown in the table 4.4 below:

Table 4.4: Mann-Kendall trend test / Two-tailed test (water demand)

Kendall's tau	0.482
S	3440.000
Var(S)	194366.667
p-value (Two-tailed)	< 0.0001
alpha	0.05

As the computed p-value is lower than the significance level alpha=0.05, we should reject the null hypothesis H_0 that there is no trend in the monthly demand for water in Kisumu City, and accept the alternative hypothesis H_1 that there is trend in the monthly demand for water in Kisumu City. These results corroborate the OLS results obtained earlier.

Using the standard normal variate;

$$\begin{aligned} Z &= \frac{S-1}{\sqrt{Var(S)}} \\ &= \frac{3440-1}{\sqrt{3440}} \\ &= \frac{3439}{58.652} \\ &= 58.634 \end{aligned} \tag{4.6}$$

From the Normal tables $Z_{\frac{\alpha}{2}} = 1.96$

Since $Z_{Calc}(58.634) > Z_{\frac{\alpha}{2}}(1.96)$, we reject the null hypothesis and conclude that there is trend in the monthly water demand for Kisumu City.

To examine the impact of misspecification of the model, the model in Equation 1 was fitted including a quadratic term. That is, $ln(y_i) = \gamma_0 + \gamma_1 t_i + \gamma_2 t_i^2 + \varepsilon$ in which

Results show that the quadratic plot yielded a MAPE value of 0.627141 which is much less that the MAPE value of the linear trend plot (1.18556). This may imply that the quadratic trend analysis fits well with the data than the linear trend analysis. Like the model without the quadratic term, the overall fit of the model with the quadratic term was good ($p < 0.05$)

Kendall's tau test for Seasonal trend

In this test, we consider the fact that the time series are seasonal with a seasonality of 12 months. The seasonal Mann-Kendall test takes into account the 12 month seasonality and tests whether there is a trend due to seasonality. The hypotheses tested were:

H_0: There is no seasonal trend in the water demand series

H_1: There is a seasonal trend in the water demand series

The results of the test are as illustrated in table 4.6 below:

Table 4.6: Seasonal Mann-Kendall Test / Period = 12 / Serial independence / Two-tailed test (water demand)

Kendall's tau	0.530
S'	286.000
p-value (Two-tailed)	< 0.0001
alpha	0.05

As the computed p-value is lower than the significance level alpha=0.05, we should reject the null hypothesis H_0 that there is no seasonal trend in the monthly demand for water in Kisumu City, and accept the alternative hypothesis Ha that there is seasonal trend in the monthly demand for water in Kisumu City. The Kendall's tau is larger when we take into account the seasonality as compared to when seasonality is not taken into account. We therefore conclude that there is a trend in the water demand time series when we take into account the seasonality.

Using the standard normal variate;

$$\begin{aligned} Z &= \frac{S' - 1}{\sqrt{Var(S')}} \\ &= \frac{286 - 1}{\sqrt{286}} \\ &= \frac{285}{16.912} \\ &= 16.852 \end{aligned} \tag{4.7}$$

From the Normal tables $Z_{\frac{\alpha}{2}} = 1.96$

Since $Z_{Calc}(16.852) > Z_{\frac{\alpha}{2}}(1.96)$, we reject the null hypothesis and conclude that there is seasonal trend in the monthly water demand for Kisumu City.

4.4 $SARIMA(p, d, q)(P, D, Q)_S$ Model For Residential Water Demand in Kisumu City

The second objective of the study sought to propose a $SARIMA(p, d, q)(P, D, Q)_S$ model that could be used to forecast residential water demand in Kisumu City. First the Kisumu City water demand data as provided by KIWASCO was tested for stationarity and autocorrelations using KPSS and ADF unit root tests, then the Box–Jenkins methodology was applied to develop the appropriate model.

4.4.1 Test for Stationary and Autocorrelation

Before testing for stationarity,time series plots were done using line plot. This is illustrated in figure 4.5.

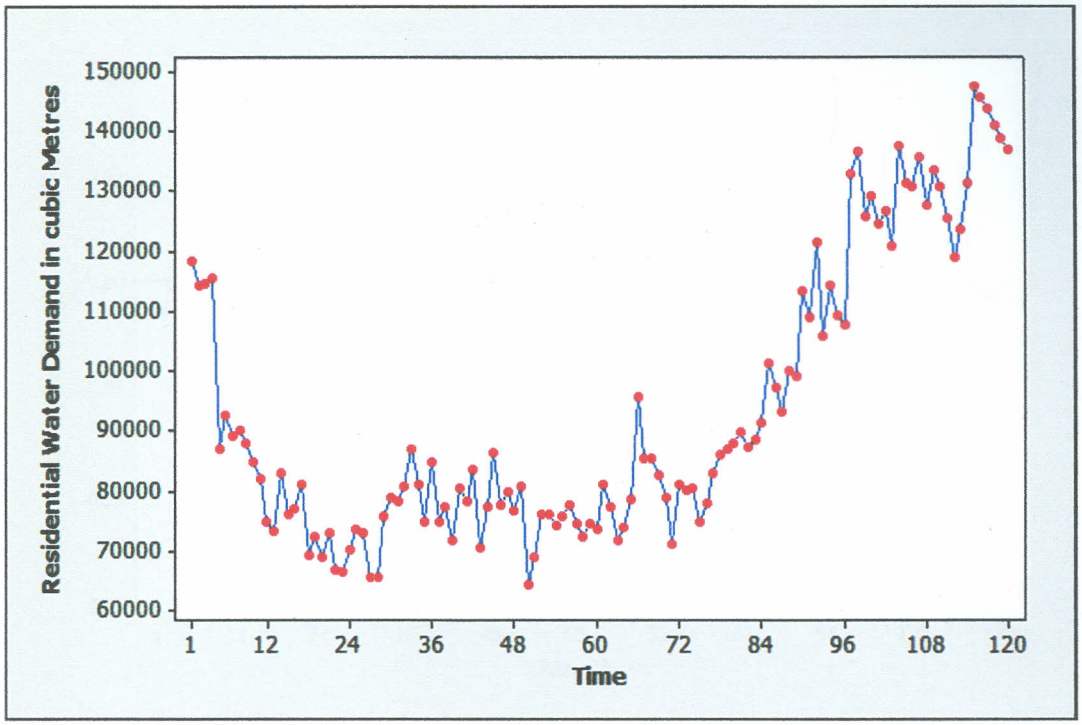


Figure 4.5: Time Series data Line plot for Water consumption in Kisumu City

The residential water demand time plot above show that the mean and variance are not constant, showing non-stationarity of the data. On average it also shows a drop in residential water demand in the first 2 years from a high of 118155 M^3 in 2004 to a low of 64341M^3 in November 2005 after which there was an increasing trend to a high of 147469M^3 in July 2013.

First differencing was done and a plot of the resulting series is obtained and is shown in the figure 4.6 below:

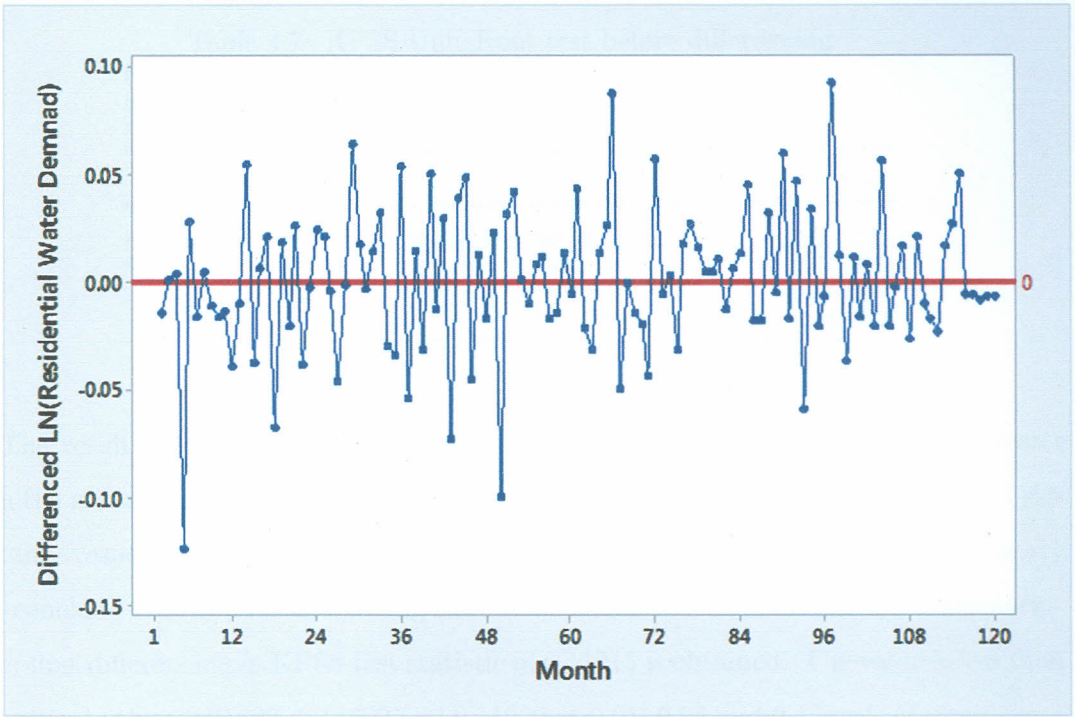


Figure 4.6: Time Series data plot for Water consumption in Kisumu City after first difference

From the above figure, it is inferred that after the first-difference, the Water demand series data does fluctuate around a constant mean hence the resulting data is stationary. Further tests were carried out using ADF and KPSS tests

4.4.2 KPSS Unit root test for Stationarity

The hypothesis to test for stationarity under KPSS is stated as follows:

$H_0 : \sigma^2 = 0$,(Stationary)

$H_a : \sigma^2 \neq 0$,(non−Stationary)

The results of the test are as shown in the table 4.7 below:

Table 4.7: KPSS Unit Root test before differencing

Null Hypothesis: WATER_VOL is stationary
Exogenous: Constant, Linear Trend
Bandwidth: 9 (Newey-West automatic) using Bartlett kernel

		LM-Stat.
Kwiatkowski-Phillips-Schmidt-Shin test statistic		0.30022...
Asymptotic critical values*:	1% level	0.21600...
	5% level	0.14600...
	10% level	0.11900...

The results of the analysis show that the KPSS test statistic of 0.30022 is greater than the critical values of 0.21600, 0.14600 and 0.11900 at 0.01, 0.05 and 0.1 levels of significance respectively. We therefore reject the null hypothesis that the data is stationary and conclude that before differencing the water demand series data are non-stationary. After first differencing, a KPSS test statistic of 0.11715 is obtained. The value is less than the critical values 0.21600, 0.14600 and 0.11900 at 0.01, 0.05 and 0.1 levels of significance respectively and hence the null hypothesis is not rejected and we conclude that the water demand series data is stationary at first difference. This is shown in the table 4.8below:

Table 4.8: KPSS Unit Root test after first difference

Null Hypothesis: D(WATER_VOL) is stationary
Exogenous: Constant, Linear Trend
Bandwidth: 23 (Newey-West automatic) using Bartlett kernel

		LM-Stat.
Kwiatkowski-Phillips-Schmidt-Shin test statistic		0.11705...
Asymptotic critical values*:	1% level	0.21600...
	5% level	0.14600...
	10% level	0.11900...

4.4.3 ADF Unit root test for Stationarity

The null hypothesis for the ADF test was that the water demand series had unit roots or the series is non-stationary. Based on the results as shown in table 4.9, we reject the null hypothesis implying that the data (before differencing) is non-stationary at 0.05 critical value.

Table 4.9: ADF Unit Root test before differencing

Null Hypothesis: WATER_VOL has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 1 (Automatic - based on SIC, maxlag=12)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-3.177329	0.0941
Test critical values:	1% level	-4.037668	
	5% level	-3.448348	
	10% level	-3.149326	

After first differencing, however, the test statistics are less than the critical values at 0.05 and hence the null hypothesis is not rejected and we conclude that the data is stationary at first difference. This is further verified by formally using the ADF tests as shown in Table 4.10.

Table 4.10: ADF Unit Root test after differencing

Null Hypothesis: D(WATER_VOL) has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 0 (Automatic - based on SIC, maxlag=12)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-14.90369	0.0000
Test critical values:	1% level	-4.037668	
	5% level	-3.448348	
	10% level	-3.149326	

Both ADF and KPSS tests confirms the non-existence of unit root under the situation where either a constant or both constant and linear trend were included in the test. Therefore, the difference order should be at least one at non-seasonal level.

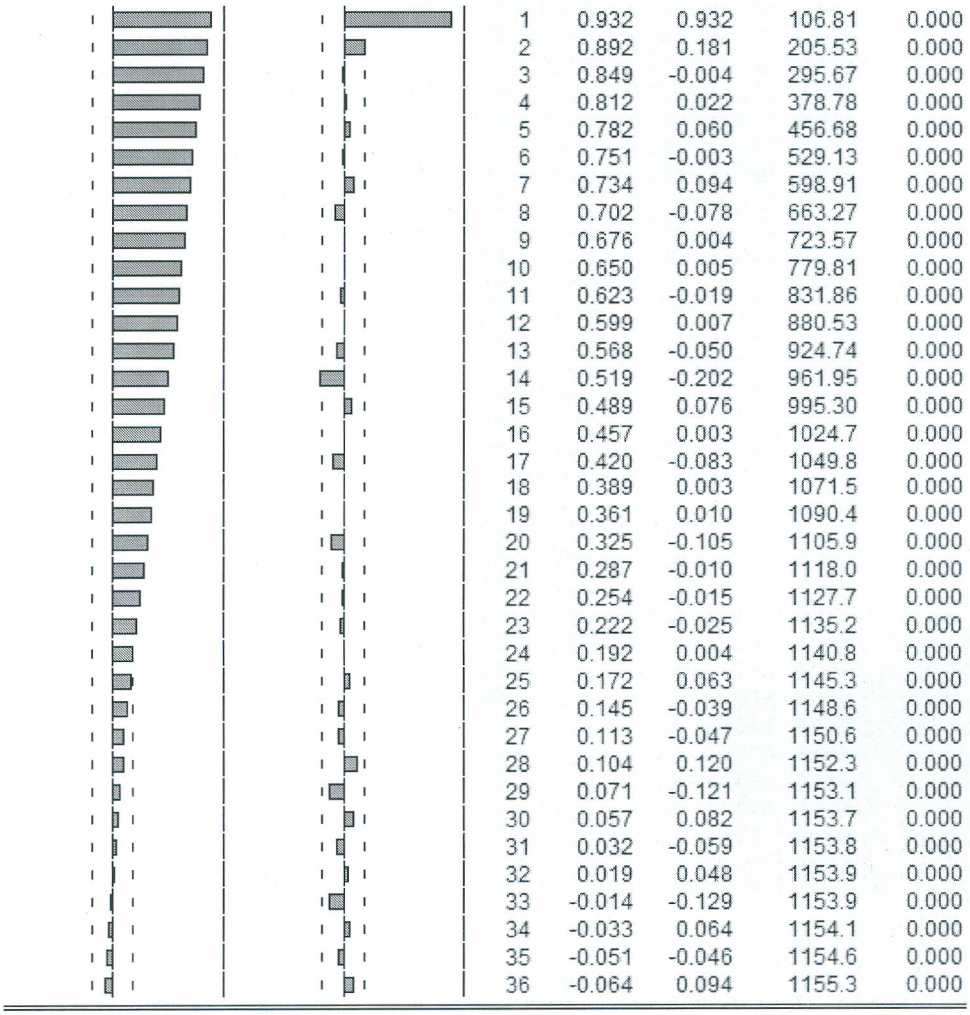
4.5 Water Demand Model Building

4.5.1 Model Identification

The results of both the KPSS and ADF unit root tests showed that the residential water demand was non-stationary before differencing but was stationary after the first differ-

ence at both seasonal and non-seasonal levels meaning that d=1 and D=1.

Date: 03/23/14 Time: 11:13
Sample: 2004M01 2013M12
Included observations: 120



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.932	0.932	106.81	0.000
		2	0.892	0.181	205.53	0.000
		3	0.849	-0.004	295.67	0.000
		4	0.812	0.022	378.78	0.000
		5	0.782	0.060	456.68	0.000
		6	0.751	-0.003	529.13	0.000
		7	0.734	0.094	598.91	0.000
		8	0.702	-0.078	663.27	0.000
		9	0.676	0.004	723.57	0.000
		10	0.650	0.005	779.81	0.000
		11	0.623	-0.019	831.86	0.000
		12	0.599	0.007	880.53	0.000
		13	0.568	-0.050	924.74	0.000
		14	0.519	-0.202	961.95	0.000
		15	0.489	0.076	995.30	0.000
		16	0.457	0.003	1024.7	0.000
		17	0.420	-0.083	1049.8	0.000
		18	0.389	0.003	1071.5	0.000
		19	0.361	0.010	1090.4	0.000
		20	0.325	-0.105	1105.9	0.000
		21	0.287	-0.010	1118.0	0.000
		22	0.254	-0.015	1127.7	0.000
		23	0.222	-0.025	1135.2	0.000
		24	0.192	0.004	1140.8	0.000
		25	0.172	0.063	1145.3	0.000
		26	0.145	-0.039	1148.6	0.000
		27	0.113	-0.047	1150.6	0.000
		28	0.104	0.120	1152.3	0.000
		29	0.071	-0.121	1153.1	0.000
		30	0.057	0.082	1153.7	0.000
		31	0.032	-0.059	1153.8	0.000
		32	0.019	0.048	1153.9	0.000
		33	-0.014	-0.129	1153.9	0.000
		34	-0.033	0.064	1154.1	0.000
		35	-0.051	-0.046	1154.6	0.000
		36	-0.064	0.094	1155.3	0.000

Figure 4.7: Collelogram of Water Demand data

Figure 4.7 above, the ACFs are suffered from linear decline and there is one significant spike of PACFs in period 1. To identify the integration order of the non-stationary time series, We take the first-difference of the series and see whether the first-difference series becomes stationary. This is illustrated in figure 4.8

Date: 03/23/14 Time: 11:15
Sample: 2004M01 2013M12
Included observations: 119

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.292	-0.292	10.423	0.001
		2	0.020	-0.072	10.471	0.005
		3	-0.086	-0.111	11.388	0.010
		4	0.017	-0.045	11.425	0.022
		5	-0.018	-0.036	11.464	0.043
		6	-0.018	-0.047	11.506	0.074
		7	0.204	0.202	16.877	0.018
		8	-0.027	0.106	16.973	0.030
		9	-0.029	0.012	17.087	0.047
		10	-0.011	0.024	17.103	0.072
		11	-0.038	-0.035	17.297	0.099
		12	0.125	0.122	19.413	0.079
		13	0.094	0.200	20.618	0.081
		14	-0.139	-0.108	23.249	0.056
		15	0.014	-0.058	23.275	0.078
		16	-0.015	-0.010	23.307	0.106
		17	0.102	0.100	24.764	0.100
		18	-0.081	-0.004	25.706	0.107
		19	0.115	0.029	27.604	0.091
		20	-0.041	-0.067	27.845	0.113
		21	-0.004	0.029	27.847	0.145
		22	-0.079	-0.039	28.780	0.151
		23	-0.012	-0.073	28.802	0.187
		24	0.053	-0.039	29.230	0.212
		25	0.065	0.039	29.871	0.229
		26	0.041	0.091	30.138	0.262
		27	-0.167	-0.104	34.514	0.152
		28	0.080	-0.013	35.519	0.155
		29	-0.107	-0.111	37.363	0.137
		30	0.077	0.033	38.317	0.142
		31	-0.038	0.004	38.554	0.165
		32	0.179	0.131	43.851	0.079
		33	-0.062	0.028	44.492	0.087
		34	-0.033	0.030	44.674	0.104
		35	-0.109	-0.081	46.707	0.089
		36	0.135	0.118	49.878	0.062

Figure 4.8: Collelogram of Water Demand data after first difference

From the above collelogram, there is one significant spike of PACFs implying presence of AR(1) process. Also, there is one significant spike of ACFs implying presence of MA(1) process. Based on the pattern, the respective values of p, d, q were determined for ARIMA part of the model given as ARIMA (1, 1, 1).

From ACF correlogram, seasonal pattern of the data is identified. As ACF is indicating seasonal pattern. Applying 12 period seasonal difference the collelogram of the

resulting series is given in the figure 4.9 below:

Date: 04/01/14 Time: 15:20
Sample: 2004M01 2013M12
Included observations: 108

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.763	0.763	64.623	0.000
		2	0.710	0.306	121.13	0.000
		3	0.598	-0.030	161.58	0.000
		4	0.520	-0.022	192.42	0.000
		5	0.479	0.085	218.85	0.000
		6	0.481	0.163	245.74	0.000
		7	0.484	0.104	273.25	0.000
		8	0.402	-0.189	292.45	0.000
		9	0.352	-0.091	307.33	0.000
		10	0.284	-0.007	317.14	0.000
		11	0.217	-0.032	322.92	0.000
		12	0.128	-0.169	324.96	0.000
		13	0.180	0.202	329.01	0.000
		14	0.124	-0.049	330.94	0.000
		15	0.130	0.010	333.11	0.000
		16	0.070	-0.132	333.75	0.000
		17	0.051	0.036	334.09	0.000
		18	-0.004	0.009	334.09	0.000
		19	-0.003	0.091	334.10	0.000
		20	-0.014	-0.103	334.12	0.000
		21	0.016	0.125	334.16	0.000
		22	0.032	0.050	334.30	0.000
		23	0.049	0.052	334.63	0.000
		24	0.043	-0.102	334.89	0.000

Figure 4.9: Collelogram of Water Demand after seasonal difference

The ACF and PACF plots indicate that the series is non-stationary with ACF showing gradual decline in values. The remaining non-stationary can be removed by further first difference. The resulting ACF and PACF plots are shown in the figure below:

Date: 04/01/14 Time: 15:23
Sample: 2004M01 2013M12
Included observations: 107

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.418	-0.418	19.251	0.000
		2	0.114	-0.074	20.702	0.000
		3	-0.094	-0.089	21.687	0.000
		4	0.072	0.006	22.277	0.000
		5	-0.218	-0.225	27.715	0.000
		6	0.040	-0.185	27.903	0.000
		7	0.205	0.199	32.799	0.000
		8	-0.047	0.141	33.054	0.000
		9	0.025	0.063	33.129	0.000
		10	0.013	0.024	33.148	0.000
		11	0.111	0.176	34.649	0.000
		12	-0.359	-0.201	50.434	0.000
		13	0.227	0.011	56.842	0.000
		14	-0.168	-0.141	60.377	0.000
		15	0.166	0.024	63.851	0.000
		16	-0.073	0.053	64.535	0.000
		17	0.176	0.055	68.529	0.000
		18	-0.168	-0.123	72.240	0.000
		19	0.055	0.047	72.645	0.000
		20	-0.103	-0.060	74.061	0.000
		21	0.034	0.010	74.220	0.000
		22	-0.037	-0.023	74.409	0.000
		23	0.095	0.087	75.656	0.000
		24	-0.099	-0.223	77.034	0.000

Figure 4.10: Collelogram of first difference of seasonally differenced Water Demand

After applying first difference on the seasonally differenced data, it can seen from the Figure 4.10 above that the ACF has a significant spike at lag 12, indicating seasonality of period 12 and a seasonal moving average component of order one. The PACF dies down and has no significant spike at lag 12 suggesting a lack seasonal autoregressive component. Therefore the inferred order of the seasonal part of the SARIMA model $(P, D, Q)_S$ is $(0, 1, 1)_{12}$.

Therefore the tentative $SARIMA(p, d, q)(P, D, Q)_S$ model will be given as:
$SARIMA(1, 1, 1)(0, 1, 1)_{12}$.

In order to make sure that the right model has been identified the following tentative models are also suggested:

(a) $SARIMA(1, 1, 1)(1, 1, 1)_{12}$

(b) $SARIMA(1, 1, 1)_{12}$

(c) $SARIMA(0,1,1)(0,1,1)_{12}$

4.5.2 Model Evaluation and Parameter Estimation

After the identification of the tentative SARIMA models, the parameters of the model are estimated using maximum likelihood estimates. The evaluation and choice of the appropriate model is based on the AIC AICc and BIC values. The model with the least values will be the best model. Table 4.11below presents the various tentative models and their corresponding AIC AICc and BIC values:

Table 4.11: AIC, AICc and BIC Values of tentative models

ARIMA MODEL	AIC	AICc	BIC
$(1,1,1)(0,1,1)_{12}$	2197.282	2197.518	2205.273
$(1,1,1)(0,0,0)_{12}$	2426.311	2426.521	2434.623
$(0,1,1)(0,,1,1)_{12}$	2199.267	2199.663	2209.921
$(1,1,1)(1,1,1)_{12}$	2201.266	2201.868	2214.580

The results show that $SARIMA(1,1,1)(0,1,1)_{12}$ with the least AIC and BIC values of 2197.282 and 2205.273 respectively and $SARIMA(0,1,1)(0,1,1)_{12}$ with AIC and BIC values of 2199.267 and 2209.921 respectively were selected as the best models for further consideration. Using the method of Maximum Likelihood, the estimated parameters of the models with their corresponding standard error are given in table 4.12 below:

Table 4.12: Estimates of the Parameters of the tentative models

MODEL	PARAMETER	COEF.	SE	t	p
$(1,1,1)(0,1,1)_{12}$	AR(1)	-0.0203	0.244	-2.08	0.039
	MA(1)	0.3746	0.229	4.35	0.024
	SMA(12)	0.8096	0.075	10.90	0.000
$(0,1,1)(0,1,1)_{12}$	MA(1)	0.3372	0.394	1.64	0.031
	SMA(12)	0.8114	0.111	10.86	0.000

Each of the SARIMA tentative model parameters are tested using t-test values and

p-values. If the p-value associated with the parameter's t-statistic is less than alpha level, we can conclude that the coefficient is significantly different from zero.

The results in table 4.12 show that all the t values for the $SARIMA(1,1,1)(0,1,1)_{12}$ model are significant since the p-values are less than the 0.05 level of significance hence it is concluded that the coefficients are significantly different from zero. This is further verified by using the t-test. Under this test we reject the null hypothesis (The coefficients are zero) if $|t| > t_{\frac{\alpha}{2}}, df = n - np$

From the t-tables, $t_{\frac{\alpha}{2}}, 120 = 1.980$ Since $|t| > t_{table}(= 1.980)$, the null hypothesis is rejected, and it is concluded that the coefficients are significantly different from zero. The coefficients of the $SARIMA(0,1,1)(0,1,1)_{12}$ model are also significant since the p-values are less than the 0.05 level of significance hence it is concluded that the coefficients are significantly different from zero. However, based on the Standard error (SE) values, it is noticed that $SARIMA(0,1,1)(0,1,1)_{12}$ has larger values as compared to $SARIMA(1,1,1)(0,1,1)_{12}$ SE values hence its estimates would be less precise.

Considering the significance of the estimated parameters for the two models, the SE values and the least AIC and BIC fit statistics, it can be established that the best fit model for residential water demand for Kisumu City is the $SARIMA(1,1,1)(0,1,1)_{12}$ The model can be written as shown below:

$$\phi_p(B)\Phi_P(B^S)\nabla^d\nabla_S^D X_t = \theta_q(B)\Theta_Q(B^S)e_t$$

$$\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D X_t = \theta_q(B)\Theta_Q(B^S)e_t \tag{4.8}$$

Substituting p=1, d=1, q=1, P=0, D=1 and Q=1 in equation (4.8) above we have;

$$\phi_1(B)\Phi_0(B^{12})\nabla^1\nabla_{12}^1 X_t = \theta_1(B)\Theta_1(B^{12})e_t$$

$$\phi_1(B)\Phi_0(B^{12})(1-B)^1(1-B^{12})^1 X_t = \theta_1(B)\Theta_1(B^{12})e_t$$

$$(1-\phi_1(B))(1-B)(1-B^{12})X_t = (1-\theta_1 B)(1-\Theta_1 B^{12})e_t$$

$$
\begin{aligned}
(1-\phi_1(B))(1-B)(X_t - X_{t-12}) &= (1-\theta_1 B)(e_t - \Theta_1 e_{t-12}) \\
(1-\phi_1(B))(X_t - X_{t-12} - BX_t + BX_{t-12}) &= e_t - \Theta_1 e_{t-12} - \theta_1 B)(e_t + \theta_1\Theta_1 B e_{t-12} \\
(1-\phi_1(B))(X_t - X_{t-12} - X_{t-1} + X_{t-13}) &= e_t - \Theta_1 e_{t-12} - \theta_1(e_{t-1} + \theta_1\Theta_1 e_{t-13} \\
(1-\phi_1(B))(X_t - X_{t-12} - X_{t-1} + X_{t-13}) &= e_t - \Theta_1 e_{t-12} - \theta_1(e_{t-1} + \theta_1\Theta_1 e_{t-13} \\
X_t - X_{t-12} - X_{t-1} + X_{t-13} - \phi_1(B)X_t + \phi_1(B)X_{t-12} + \phi_1(B)X_{t-1} - \phi_1(B)X_{t-13} &= \\
e_t - \Theta_1 e_{t-12} - \theta_1(e_{t-1} + \theta_1\Theta_1 e_{t-13} & \\
X_t - X_{t-12} - X_{t-1} + X_{t-13} - \phi_1 X_{t-1} + \phi_1 X_{t-13} + \phi_1 X_{t-2} - \phi_1 X_{t-14} &= \\
e_t - \Theta_1 e_{t-12} - \theta_1(e_{t-1} + \theta_1\Theta_1 e_{t-13} & \\
X_t - X_{t-1} - \phi_1 X_{t-1} + \phi_1 X_{t-2} - X_{t-12} + X_{t-13} + \phi_1 X_{t-13} - \phi_1 X_{t-14} &= \\
e_t - \Theta_1 e_{t-12} - \theta_1(e_{t-1} + \theta_1\Theta_1 e_{t-13} & \\
X_t - (1+\phi_1)X_{t-1} + \phi_1 X_{t-2} - X_{t-12} + (1+\phi_1)X_{t-13} - \phi_1 X_{t-14} &= \\
e_t - \Theta_1 e_{t-12} - \theta_1 e_{t-1} + \theta_1\Theta_1 e_{t-13} &
\end{aligned}
\tag{4.9}
$$

Making X_t the subject we have;

$$
\begin{aligned}
X_t = (1+\phi_1)X_{t-1} - \phi_1 X_{t-2} + X_{t-12} - (1+\phi_1)X_{t-13} + \phi_1 X_{t-14} + & \\
e_t - \Theta_1 e_{t-12} - \theta_1 e_{t-1} + \theta_1\Theta_1 e_{t-13} &
\end{aligned}
\tag{4.10}
$$

Since logarithm values were used in the modelling, the $SARIMA(1,1,1)(0,1,1)_{12}$ model becomes

$$
\begin{aligned}
ln(X_t) = (1+\phi_1)ln(X_{t-1}) - \phi_1 ln(X_{t-2}) + ln(X_{t-12}) - (1+\phi_1)ln(X_{t-13}) + & \\
\phi_1 ln(X_{t-14}) + e_t - \Theta_1 e_{t-12} - \theta_1 e_{t-1} + \theta_1\Theta_1 e_{t-13} &
\end{aligned}
\tag{4.11}
$$

But,

$\phi_1 = -0.0203$

$\theta_1 = 0.3746$

$\Theta_1 = 0.8096$

Substituting these values in equation above we have;

$$\begin{aligned} ln(X_t) &= 0.9797ln(X_{t-1}) + 0.0203ln(X_{t-2}) + ln(X_{t-12}) - 0.9797ln(X_{t-13}) - \\ &\quad 0.0203ln(X_{t-14}) + e_t - 0.8096e_{t-12} - 0.3746e_{t-1} + 0.3033e_{t-13} \\ &= 0.9797ln(X_{t-1}) + 0.0203ln(X_{t-2}) + ln(X_{t-12}) - 0.9797ln(X_{t-13}) - \\ &\quad 0.0203ln(X_{t-14}) + e_t - 0.8096e_{t-12} + 0.0.3746e_{t-1} + 0.3033e_{t-13} \\ &= ln(X_{t-12}) + \langle 0.9797ln(X_{t-1}) + 0.0203ln(X_{t-2}) - 0.9797ln(X_{t-13}) - \\ &\quad 0.0203ln(X_{t-14})\rangle + e_t - 0.3746e_{t-1} - 0.8096e_{t-12} + 0.3033e_{t-13} \\ &= ln(X_{t-12}) + ln\langle\frac{X_{t-1}^{0.9797} * X_{t-2}^{0.0203}}{X_{t-13}^{0.9797} * X_{t-14}^{0.0203}}\rangle + \langle e_t - 0.3746e_{t-1} - 0.8096e_{t-12} + \\ &\quad 0.3033e_{t-13}\rangle \end{aligned} \tag{4.12}$$

This means that holding all factors constant, this month residential water demand is the sum of

(i) the value of the time series in the same month of the previous year,

(ii) a trend component determined by the difference between the sum of previous month's value and the previous two month's value and the sum of last year's previous month's value and last year's previous two month's value;

(iii) the effects of the residual terms of period t, t-1, t-12 and t-13 on the forecast.

4.5.3 Model Adequacy Checking

This aims at examining the accuracy of the chosen tentative model in ensuring that the modelling assumptions are satisfied.

First the Ljung-Box (Q) test was used for testing white noise residual. The Hypotheses to be tested were:

H_0:Residuals are white noise

H_a:Residuals are not white noise

H_0 is rejected if the p-value of the Q-statistic is less than the 0.05 level of significance. The ACF and PACF at some lags together with Q statistics for the Box Ljung test of the residuals are shown in the figure 4.11below:

Date: 04/02/14 Time: 14:19
Sample: 2004M01 2013M12
Included observations: 119

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.004	0.004	0.0017	0.967
		2	0.014	0.014	0.0255	0.987
		3	-0.073	-0.073	0.6812	0.878
		4	0.076	0.077	1.4086	0.843
		5	0.016	0.017	1.4411	0.920
		6	0.026	0.018	1.5240	0.958
		7	0.238	0.251	8.7882	0.268
		8	0.054	0.051	9.1733	0.328
		9	-0.018	-0.025	9.2168	0.418
		10	-0.044	-0.010	9.4737	0.488
		11	0.046	0.015	9.7610	0.552
		12	0.157	0.148	13.088	0.363
		13	0.144	0.147	15.893	0.255
		14	-0.099	-0.168	17.239	0.244
		15	0.037	0.028	17.432	0.294
		16	0.032	0.054	17.579	0.349
		17	0.065	0.040	18.183	0.377
		18	-0.096	-0.100	19.497	0.362
		19	0.096	0.010	20.814	0.347
		20	0.001	-0.084	20.814	0.408
		21	-0.003	0.042	20.815	0.470
		22	-0.070	-0.045	21.552	0.487
		23	0.007	-0.054	21.559	0.547
		24	0.062	0.008	22.136	0.571

Figure 4.11: Collelogram of Residuals of $SARIMA(1,1,1)(0,1,1)_{12}$

The results show that none of the Q statistics is statistically significant,i.e. their p-values all exceed 0.05 for all lag orders. This indicates that there is no significant departure from white noise for the residuals. Therefore we fail to reject H_0 and conclude that the standardized residuals follow a white noise process with a mean of zero and constant variance. This also indicates the absence of autocorrelation. This is based on the recommendation by Wei[?] that If the Q statistic is significant then the model is not adequate, and if the Q statistic is not significant then the fitted ARIMA model is appropriate. Any significant autocorrelation may be an indication of misspecification.

Also, the ACF and PACF plots of the residuals show that the ACF of the residuals immediately die out from lag one (1), which means the residual are white noise. The zero

mean and constant variance assumption is further illustrated in the following time series plot of the standardized residuals in figure 4.12 below:

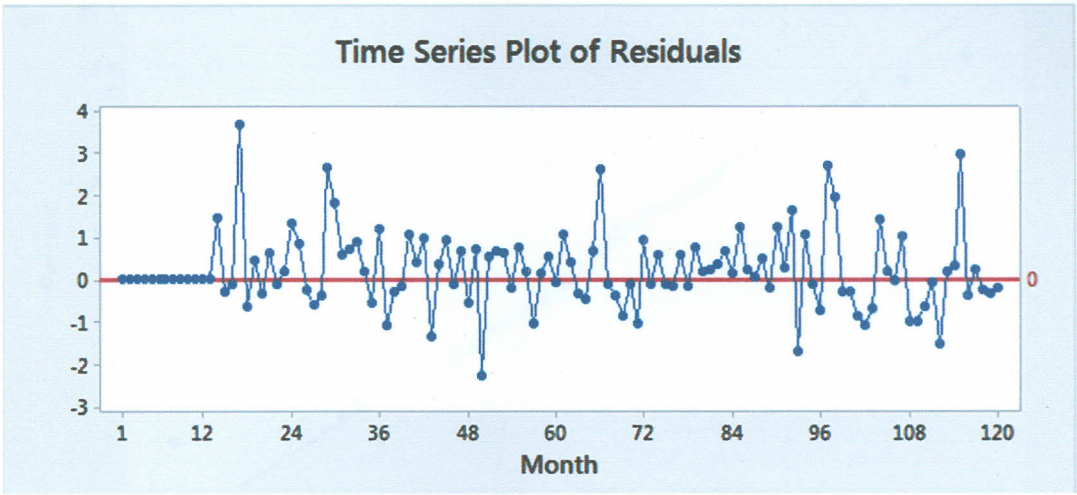


Figure 4.12: Time Series plot of the standardized Residuals of $SARIMA(1,1,1)(0,1,1)_{12}$

To test normality of the residuals, the normal Probability plot of standardized residuals was used. The results as shown in figure 4.13 below shows that the normal probability plot of the residuals is approximately linear supporting the condition that the error terms are normally distributed. This is because the data fall close to the line representing a normal distribution. This implies that for short term forecasting, SARIMA model can reproduce the details of the original series

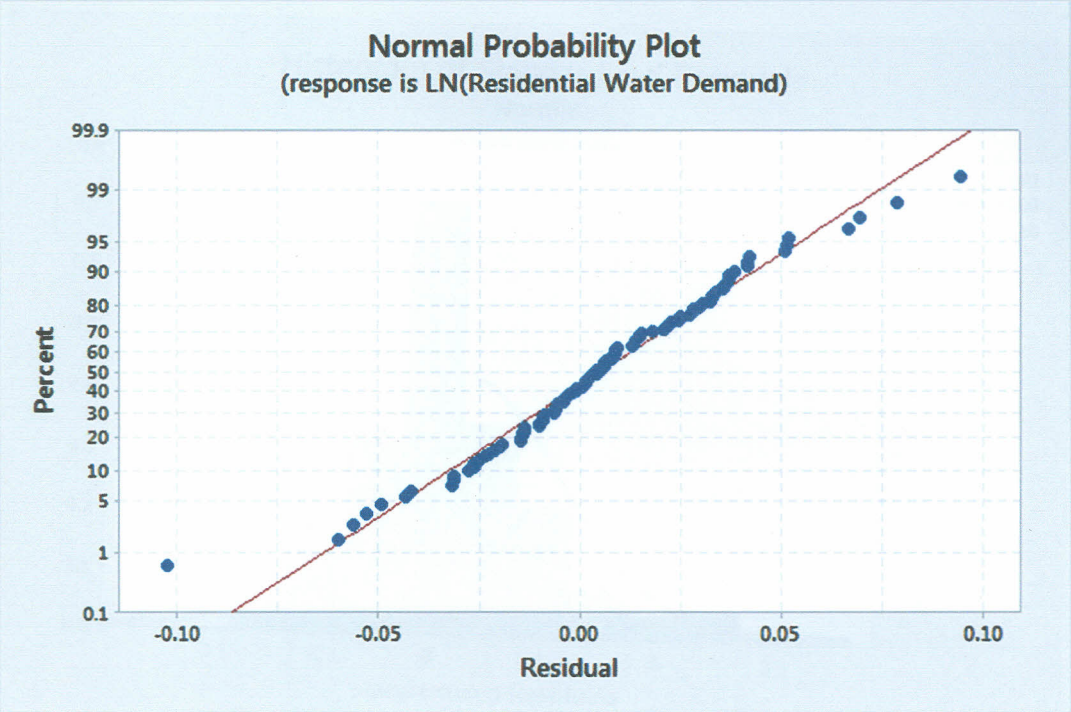


Figure 4.13: Normal Probability plot of the Residuals of $SARIMA(1,1,1)(0,1,1)_{12}$

Also the histogram plot of standardized residuals in figure ?? below show that the standardized residuals are normally distributed with $Mean \simeq 0$ and $Variance \simeq 1$

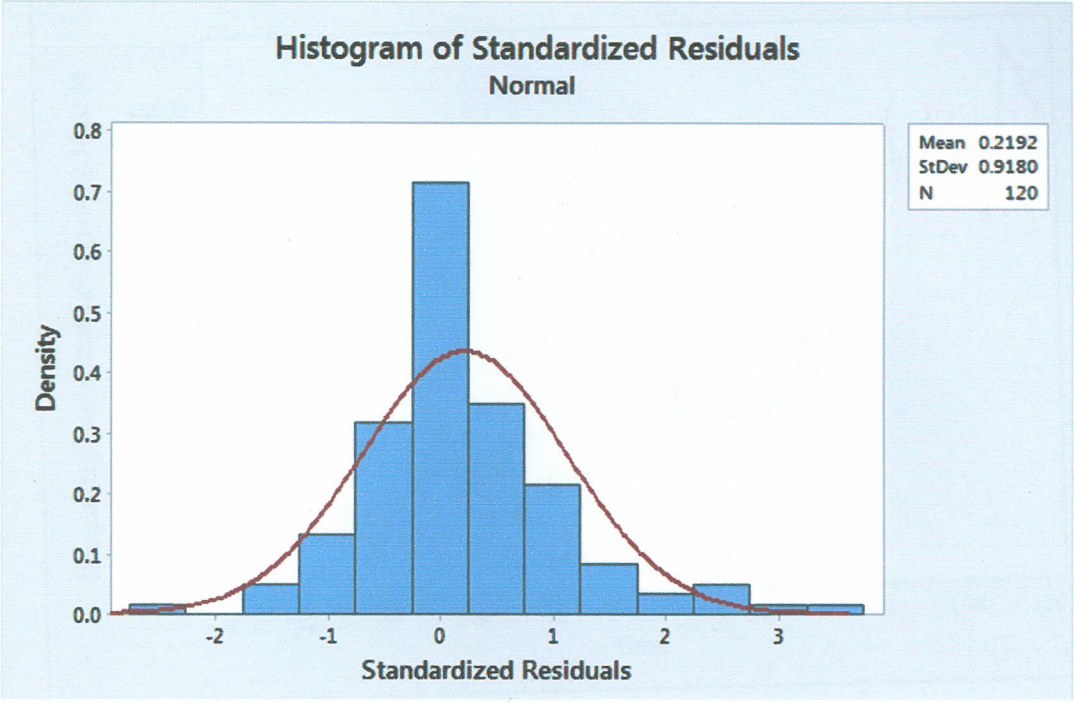


Figure 4.14: Histogram plot of the Standardized Residuals of $SARIMA(1,1,1)(0,1,1)_{12}$

Finally, the actual data was graphically compared with model data so as to assess the agreement between their plots. Figure 4.15 below shows a very close agreement between the fitted model and the actual data.

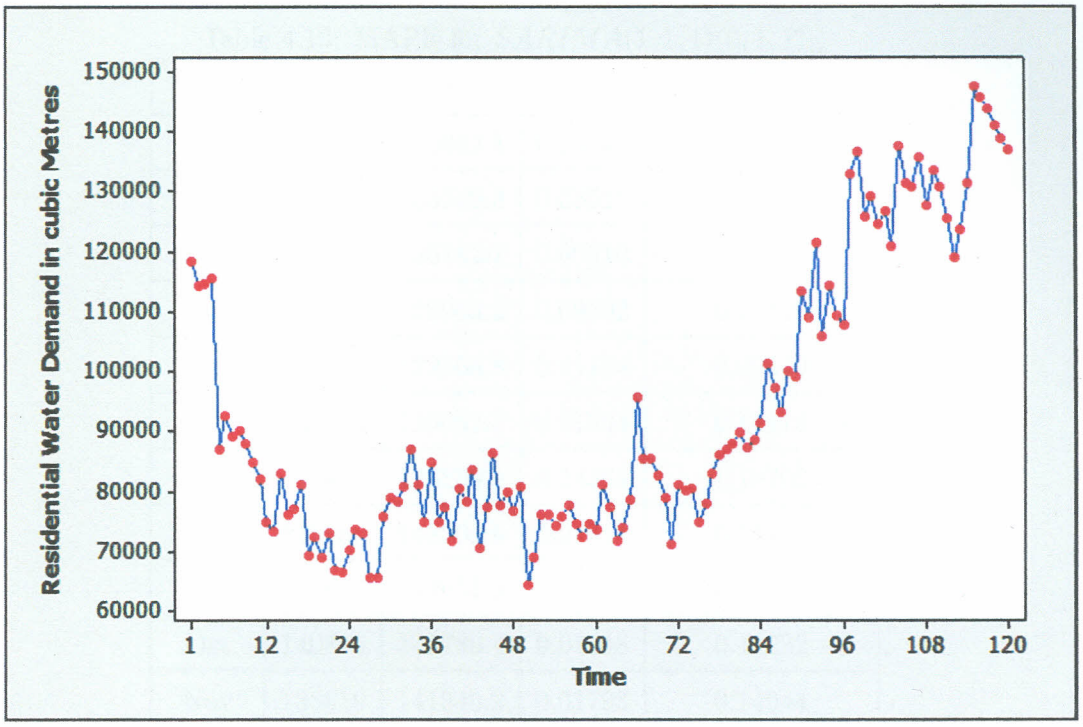


Figure 4.5: Time Series data Line plot for Water consumption in Kisumu City

The residential water demand time plot above show that the mean and variance are not constant, showing non-stationarity of the data. On average it also shows a drop in residential water demand in the first 2 years from a high of 118155 M^3 in 2004 to a low of 64341M^3 in November 2005 after which there was an increasing trend to a high of 147469M^3 in July 2013.

First differencing was done and a plot of the resulting series is obtained and is shown in the figure 4.6 below:

Table 4.13: MAPE for $SARIMA(1,1,1)(0,1,1)_{12}$

Month	A_t	F_t	$\frac{\lvert A_t - F_t \rvert}{A_t}$	$\frac{\lvert A_t - F_t \rvert}{A_t} * \frac{100}{12}$
Jan	133653	140662.1	0.05245	0.43704
Feb	130646	135240.3	0.03517	0.29308
Mar	125556	126142.7	0.00510	0.04249
Apr	118962	129963.2	0.09292	0.77429
May	123569	122168.8	0.01134	0.09449
Jun	131405	129092.7	0.01694	0.14114
Jul	147469	126259.5	0.14364	1.19703
Aug	145580	148203.5	0.01817	0.15138
Sep	143647	142051.0	0.01218	0.10149
Oct	140854	142789.4	0.01348	0.11232
Nov	138819	141345.2	0.01793	0.14944
Dec	136800	138362.5	0.01148	0.09567
-	-	-	-	$MAPE = 3.590$

Also the MAPE of the OLS equation was as computed in the table below:

Table 4.14: MAPE for the OLS equation

Month	A_t	F_t	$\frac{\lvert A_t - F_t \rvert}{A_t} * \frac{100}{12}$
Jan	133653	113271	1.271
Feb	130646	113792	1.075
Mar	125556	114314	0.746
Apr	118962	114840	0.289
May	123569	115367	0.553
Jun	131405	115897	0.983
Jul	147469	116430	1.754
Aug	145580	116965	1.638
Sep	143647	117502	1.517
Oct	140854	118042	1.350
Nov	138819	118585	1.215
Dec	136800	119130	1.076
-	-	-	$MAPE = 12.196$

Comparatively, the MAPE value for $SARIMA(1,1,1)(0,1,1)_{12}$ which was 3.590 was less than the MAPE value of the OLS equation $Ln(\hat{Y}_i) = 4.8371 + 0.001991t_i$ which was 12.196. Based on the MAPE value it is concluded that the $SARIMA(1,1,1)(0,1,1)_{12}$ has a better forecasting performance.

Root Mean Square Error(RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i} - X_{model,i})^2}{n}} \tag{4.14}$$

Where;

$X_{obs,i}$ are the observed values at time i

$X_{model,i}$ are the model values at time i

The result of inspection is computed as shown in the table below:

Table 4.15: RMSE for $SARIMA(1,1,1)(0,1,1)_{12}$

Month	$X_{obs,i}$	$X_{model,i}$	$(X_{obs,i} - X_{model,i})^2$
Jan	133653	140662.1	49132253
Feb	130646	135240.3	21112411
Mar	125556	126142.7	409837.7
Apr	118962	129963.2	122177479
May	123569	122168.8	1963223.4
Jun	131405	129092.7	4953401.2
Jul	147469	126259.5	448720898
Aug	145580	148203.5	6993950.4
Sep	143647	142051.0	3060645.6
Oct	140854	142789.4	3604419.8
Nov	138819	141345.2	6196958.2
Dec	136800	138362.5	2466527.4
-	-	-	$\frac{\sum_{i=1}^{n}(X_{obs,i}-X_{model,i})^2}{12} = 55899334$
-	-	-	$RMSE = 7476.59$

The RMSE of the OLS equation is as given below:

Table 4.16: RMSE for the OLS equation

Month	$X_{obs,i}$	$X_{model,i}$	$(X_{obs,i} - X_{model,i})^2$
Jan	133653	113271	415423102.8
Feb	130646	113792	284072601
Mar	125556	114314	126373222.7
Apr	118962	114840	16993464.77
May	123569	115367	67266700.86
Jun	131405	115897	240483119
Jul	147469	116430	963417814.5
Aug	145580	116965	818817071.4
Sep	143647	117502	683536391.4
Oct	140854	118042	520369477.2
Nov	138819	118585	409423131
Dec	136800	119130	312239968.5
-	-	-	$\frac{\sum_{i=1}^{n}(X_{obs,i}-X_{model,i})^2}{12} = 4858416065$
-	-	-	$RMSE = 20121.33$

Comparatively, the RMSE value for $SARIMA(1,1,1)(0,1,1)_{12}$ which was 7476.59 was less than the RMSE value of the OLS equation $Ln(\hat{Y}_i) = 4.8371 + 0.001991t_i$ which was 20121.33 hence the conclusion that the $SARIMA(1,1,1)(0,1,1)_{12}$ has a better forecasting performance.

Table 4.17below shows the summary results of model comparison for MAPE and RMSE for each model.

Table 4.17: Summary Accuracy Measures of the Models

Perfomance Evaluation procedure	OLS model	SARIMA model
MAPE	12.196	3.590
RMSE	20121.33	7476.59

The analysis show that $SARIMA(1,1,1)(0,1,1)_{12}$ has less MAPE and RMSE values of 3.590 and 7476.59 respectively. Based on these results it is concluded that it is a better

forecasting model hence it is proposed for the forecasting of residential water demand for Kisumu City.

4.6 Residential Water Use Forecasting for 2014

The principal objective of time series modelling and analysis is forecasting. Therefore, the third objective of the study sought to forecast residential water demand for Kisumu city in the twelve months of 2014

The $SARIMA(1,1,1)(0,1,1)_{12}$ was used to generate the forecast of residential water demand for the period January 2014 to December 2014. To forecast one period ahead that is X_{t+1} the forecast equation is given by:

$$ln(\hat{X}_{t+1}) = ln(X_{t-11}) + ln\langle\frac{X_t^{0.9797} * X_{t-1}^{0.0203}}{X_{t-12}^{0.9797} * X_{t-13}^{0.0203}}\rangle + \langle\hat{e}_{t+1} - 0.3746\hat{e}_t - 0.8096\hat{e}_{t-11} + 0.3033\hat{e}_{t-12}\rangle \tag{4.15}$$

The term e_{t+1} is not known because the expected value of future random errors is taken to be zero. for instance, to forecast the residential water demand for the period 121 which is January 2014 the equation becomes:

$$ln(\hat{X}_{121}) = ln(X_{109}) + ln\langle\frac{X_{120}^{0.9797} * X_{119}^{0.0203}}{X_{108}^{0.9797} * X_{107}^{0.0203}}\rangle + \langle\hat{e}_{121} - 0.3746\hat{e}_{120} - 0.8096\hat{e}_{109} + 0.3033\hat{e}_{108}\rangle \tag{4.16}$$

$\hat{e}_{121} = 0$, $\hat{e}_{120} = X_{120} - \hat{X}_{120}$, $\hat{e}_{109} = X_{109} - \hat{X}_{109}$, $\hat{e}_{108} = X_{108} - \hat{X}_{108}$

The forecast values as well as a 95 percent confidence intervals obtained using Minitab are presented in Table 4.18 below.

Table 4.18: Residential Water Demand for $SARIMA(1,1,1)(0,1,1)_{12}$

Month	SARIMA Forecast	Observed value	95 LowerCI	95 Upper CI
JAN04	150169	158117	130860	164224
FEB04	147370	154497	126972	164926
MAR04	140145		119122	161167
APR04	145885		118117	163888
MAY04	145704		117269	164484
JUN04	153571		122317	174752
JUL04	150567		122833	178300
AUG04	155995		126643	184987
SEP04	152005		121464	182546
OCT04	150964		119111	182816
NOV04	155172		117060	183283
DEC04	149090		113359	184820

Comparing the forecast water demand for January -February 2014 with the observed water demand, it is established that the forecast values are close to the true values and do lie within the confidence intervals hence it is concluded that $SARIMA(1,1,1)(0,1,1)_{12}$ is adequate to be used to forecast monthly residential water demand in Kisumu City. It is noticeable that the confidence interval becomes wider as the number of forecasts increase. The wider confidence limit gives an indication of high stochasticity in the data.

Chapter 5

Summary, Conclusions and Recommendations

5.1 Summary of Findings and Conclusions

The first objective of the study sought to analyse the of residential water demand for Kisumu City for the years 2004 to 2014. The study employed Ordinary Least Squares (OLS) and Kendall's tau tests methods to test for trend in the data. The OLS approach established a positive significant trend represented by the equation $Ln(\hat{Y}_i) = 4.8371 + 0.001991t_i$. The F-test based on ANOVA showed that the coefficient $\gamma_1 = 0.001991$, was statistically and significantly not equal to zero at 0.05 level of significance. Also the Kendall's tau test established that $Z_{Calc}(58.634) > Z_{\frac{\alpha}{2}}(1.96)$, and concluded that there was significant positive trend in the monthly water demand for Kisumu City. Both test statistics gave the same conclusion of the existence of a meaningful statistical trend in the amount of residential water demanded in Kisumu City over the last 10 years. The study therefore concludes that the residential water demand in Kisumu City will significantly continue to increase in the future.

The second objective of the study sought to propose a SARIMA model that could be used to forecast residential water demand in Kisumu City. Following the Box-Jenkins approach and based on minimum AIC, AICc and BIC values, the best-fitted SARIMA models were $SARIMA(1,1,1)(0,1,1)_{12}$ and $SARIMA(0,1,1)(0,1,1)_{12}$ models. After the es-

timation of the parameters of selected model, a series of diagnostic and forecast accuracy test were performed. Having satisfied all the model assumptions, $SARIMA(1,1,1)(0,1,1)_{12}$ model was judged to be the best model for forecasting. Model validation based on MAPE and RMSE established that, compared to the OLS equation, the $SARIMA(1,1,1)(0,1,1)_{12}$ had better forecasting performance and hence proposed for the forecasting of residential water demand in Kisumu City.

The third objective of the study sought to forecast residential water demand for the 12 months of 2014. The Forecasting results in general revealed an increasing pattern of residential water demand over the forecast period. The study established that the forecast values were close to the true values and were within the confidence intervals hence concluded that $SARIMA(1,1,1)(0,1,1)_{12}$ was adequate and could be used to forecast monthly residential water demand in Kisumu City. However, it was observed that there were wider confidence limits which could be an indicator of high stochasticity in the water demand data.

5.2 Recommendations

5.2.1 Recommendations for policy and practice

The study demonstrates that the Box – Jenkins approach used in this study could be useful for the modelling and forecasting of residential water demand. The results of this study would be useful to water companies more so KIWASCO in exploring realistic decision making scenarios for designing effective water demand management policies. Such policies would ensure that short term future water demand is met. Second, Water Resources Management Authority (WRMA) which is a state corporation charged with the responsibility of being the lead agency in water resources management could apply the study findings to develop forecasts and projections of water use in Kenya hence provide better advice on water resources development.

5.2.2 Suggestions for future research

The accuracy of forecast future values is the core point for every forecaster. This is because the forecast values will affect the quality of the policies implemented based on this forecast. With this motive, it is therefore recommended that future research will be helpful to assess the performance of the model in terms of forecast precision as compared to other time series models.

(a) The study recommends a similar study using other water companies' data so as to corroborate the study findings.

(b) Since the current study was based on aggregated data, it is recommended that future research that is based on household level data be carried out.

(c) The study also recommends that future research aimed at modelling residential water demand apply other methodologies such as Artificial Neural Networks and Markov models. These will be helpful in assessing the forecasting accuracy of the developed model

(d) The study finally recommends future research that will estimate the parameters of the three-parameter lognormal distribution for residential water demand.

References

[1] Bithas, K. and Stoforos, C. (2006). *Estimating Urban Residential Water demand and forecasting water demand for Athens Metropolitan.* South Eastern Europe Journal of Economics (2006)47-59

[2] Box, G. M., and George, L. E. (1987). *On a measure of a Lack of Fit in Time Series Models.* Biometrika(65), 297-303 1987.

[3] Box,G. E. P.and Jenkins, G.M., (1976). *Time Series Analysis: Forecasting and Control* Holden day

[4] Buckman .A. and Mintah .E.,(2013) *An Autoregressive Moving Average (ARIMA) model for Ghana's Inflation (1985 - 2011)*, Mathematical theory and Modelling, Vol. 3, No. 3.

[5] Caldwell J.G., 2006. *The Box-Jenkins Forecasting Technique Posted at Internet.* websites http://www.foundationwebsite.org.

[6] Cryer J.D and Kung-Sik C. (2008). *Time Series Analysis with Applications in R. 2Ed.* Springer Science +Business Media, LLC, NY, USA.

[7] Dickey, D.A. and Fuller, W.A. (1981). *Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root.* Econometrica, 49(4), 1057-1072.

[8] Galiani S, Gertler P, Schargrodsky E (2005). *'water for Life: The Impact of the Privatization of Water Services on Child Mortality'*, J. Polit. Econ. 113(1): 83-120.

[9] Gupta, S., P.(2005). *'Statistical Methods*, Sultan Chand and Sons Educational Publishers, New Delhi.

[10] Habib, M., Sadegh, R., Saralees, N. and Mahsa, E. (2013). *Estimation of water demand in Iran based on SARIMA models.* Journal of Environmental Modelling and Assessment. Vol 18(5). Springer Journals 2013

[11] Halim, S. and Bisono, I.N. (2008). *Automatic Seasonal Autoregressive Moving Average Models and Unit Root Test Detection.* International Journal of Management Science and Engineering Management, 3(4): 266-274

[12] Jorge C., 2007 *Forecasting water consumption in Spainusing univariate time series models.* Online at http://mpra.ub.uni-muenchen.de/6610/ MPRA Paper No. 6610, posted 7. January 2008 04:34 UTC

[13] Kahya, E. and S. Kalayci, 2004. *Trend analysis of stream flow in Turkey.* J. Hydrol., 289: 128-144. DOI:10.1016/j.jhydrol.2003.11.006.

[14] Kihoro, J.M., Otieno, R.O., and Wafula, C. (2004). *Seasonal Time Series Forecasting: A Comparactive Study of ARIMA and ANN Models.* African Journal of Science and Technology, 5(2): 41-49

[15] KIWASCO (2007). *Kisumu Water and Sewerage Company Strategic Plan 2007-2012.*

[16] Kleiber, C. and Zeileis, A. (2008) *Applied Econometrics with R.* Springer Science +Business Media, LLC, NY, USA.

[17] Maamar S.(2013). *ANN versus SARIMA models in forecasting residential water consumption in Tunisia.*Journal of Water, Sanitation and Hygiene for Development, 2013

[18] Mahsin, M., Yesmin, A., and Monira, B. (2012) *Modelling Rainfall in Dhaka Division of Bangladesh Using Time Series* Journal of Mathematical Modelling and Application 2012, Vol. 1, No.5, 67-73

[19] Maoulidi, M. (2010): *A Water and Sanitation Needs Assessment for Kisumu City* MCI Social Sector Working Paper series NO. 12/2010

[20] Martinez-Espineira R. (2005): *An Estimation of Residential Water Demand Using Co-Integration and Error Correction Techniques.* Forthcoming in: Journal of Applied Economics

[21] Montgomery, D. C., Jennings, C. L.and Kulahci, M.(2008). *Introduction to Time Series Analysis and Forecasting.* John Wiley and Sons, Inc.

[22] Montgomery D.C. (2009). *"Design and analysis of experiments".*John Wiley and Sona, Inc., 7th edition, 2009.

[23] Onoz, B. and Bayazit, M. (2003). *The power of statistical tests for trend detection.* Turkish J. Eng. Environ. Sci., 27: 247-251.

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

[25] Pankratz, A. (1983) *Forecasting with Univariate Box-Jenkins Models: Concepts and Cases.* John Wiley and Sons. Inc. USA

[26] Racine, J.S. (2008). *Nonparametric econometrics: A primer.* Foundat. Trends Econ., 3: 1-88. DOI: 10.1561/0800000009

[27] Ruey, S. T. (2005). *Analysis of Financial Time Series.* Wiley Interscience. A John Wiley and Sons Publication

[28] Schleich J., Hillenbrand T., (2009) *Determinants of residential water demand in Germany.* Ecological Economics, Volume 68, Issue 6, 15 April 2009, Pages 1756-1769, ISSN 0921-8009, http://dx.doi.org/10.1016/j.ecolecon.2008.11.012. http://www.sciencedirect.com/science/article/pii/

[29] Smaoui, N.; BuHamra, S.; Gabr, M.,(2002) *Ä combination of Box-Jenkins analysis and neural networks to model and predict water consumption in Kuwait,"* Neural Networks, 2002. IJCNN '02. Proceedings of the 2002 International Joint Conference on , vol.2, no., pp.1678,1683, 2002. doi: 10.1109/IJCNN.2002.1007770:

[30] Stake, R. E. (2008). *Qualitative case studies. In N. K. Denzin and Y. S. Lincoln (Eds.), Strategies of qualitative inquiry* (3rd ed., pp. 119150). Thousand Oaks, CA: Sage.

[31] UNDP (2006). *Human Development Report 2006. Beyond Scarcity: Power, Poverty, and the Global Water Crisis.* London : Palgrave Macmillan for United Nations Development Programme.

[32] Wagah, G., Onyango, G. and Kibwage, J. (2010). *Accessibility of water services in Kisumu Municipality Kenya.* Journal of Geography and Regional Planning. Vol. 3(4) pp 114-125.

[33] Wooldridge, J.M., 2001. *Applications of generalized method of moments estimation.* J. Econ. Perspectives, 15: 87-100.

[34] World Bank (2003) *Water Resources Strategy.* World Bank, Washington, D.C.

[35] World Bank (2005) *Water for the Urban Poor: Water Markets, Household Demand, and Service Preferences in Kenya.* Water Supply and Sanitation Sector Board p. 5

[36] Worthington, A., C. (2010). *Commercial and Industrial water demand estimation: Theoretical and Methodological guidelines for applied economics research.* Estudios de Economa Aplicada, 2010: 237-258 Vol. 28-2.

[37] Xinming, M., Dale,W., and Briscoe, J. (2010). *Modelling Village Water Demand. A discrete choice approach* Water resources Research, 1990: Vol. 26-4, 521-529.

[38] Yaw, M., O.(2004). *Household Water Security and Water demand in Volta basin of Ghana* Unpublished Phd Thesis, 2004.