ANALYSIS OF SEASONAL TIME SERIES WITH MISSING OBSERVATIONS:
A case of harvesting of fish in Lake Victoria Kenya.

by

OTWANDE ANDREA

A project report submitted in partial fulfilment
of the requirements for the degree of Master of Science in Applied Statistics

School of Mathematics, Statistics and Actuarial Science

MASENO UNIVERSITY

©2013.

Abstract

Time series is a measured observation recorded with time. This statistical procedure is applicable in many fields of study including engineering and economics. The process of collecting data sometimes faces a lot of challenges that may arise due to defective working tools, misplaced or lost records and errors that are prone to occur. These problems can be addressed by estimating the missing values so as to enable one to proceed with the analysis and forecasting. The most commonly used approaches include the use of autoregressive-moving average models developed by Box Jenkins, use of extrapolation or interpolation under regression analysis and use of state space models where data is considered as a combination of level, trend and seasonal components. This project intends to use the most appropriate method of estimating missing values by using the direct method of imputation. Incomplete secondary data obtained from the Ministry of fisheries and Development, together with the Kenya Marine and Fisheries Research Institute are to be used to estimate the gap left just before, during and immediately after the post election violence of the year 2007/2008, a time when data could not be obtained and/or recorded. The original time series data when analysed produced a SARIMA model $(0, 1, 1)(2, 0, 0)_{12}$ as the best candidate for the lower segment. SARIMA $(0, 1, 2)(0, 0, 1)_{12}$ was produced for the upper segment using autoarima function in R package. The missing data were estimated using forecast from the lower segment which was extended to the in sample forecast in the upper segment. The regression test between predicted and the original values in upper segment proves strong positive relationship indicating high level of accuracy on predictability of the model used.

Chapter 1

Introduction

1.1 Background of the Study

Fishing industry is one of the major economic activities for the people living around Lake Victoria.They harvest fish on daily basis and sell the catch to traders to export and / or to the local consumers when fish is still fresh. Different fish species are harvested in the lake with fishing of Nile perch (Lates Niloticus) leading in priority due to its market value and nutrition standards. Data about the catch are usually recorded by the Government of Kenya through the Ministry of Livestock and Fisheries Development at various beaches around the lake shore together with the Kenya Marine and Fisheries Research Institute (KMFRI). In the year 2007 ,just before 2007 general election there was noted general laxity on fishing activity and data collection. This was followed by the post election violence where no data could be collected and even in some areas the records were lost to fire. This left a seasonal gap of missing observations for a period of three months, a problem that forms the basis of this project work," Analysis of time series with missing observations ". The problem of missing values in time series is common in data collection. One of the main objectives of time series analysis is to fill the missing observations so as to enable analysis and forecasting be done. This can be possible if a suitable model that fits the data available is used appropriately such that it can be extended to cover the missing gaps. A lot of research work has been done in this area with the stochastic models developed by Box and Jenkins widely applied in adjusting the estimates both directly and indirectly and eventual forecasting. This is because it provides a common frame work for time series forecasting that can cope with non stationary series by use of differencing technique. The technique derives forecast of time series on the basis of historical behavior of the series itself hence can use the statistical concepts and principles that can model a wide range of time series behavior. This project therefore intends to identify the model that can fit the collected data prior and after the gap and then apply the most appropriate method of adjusting the estimates and eventually forecast.

1.2 Statement of the problem

The economic activity of harvesting Nile perch in Lake Victoria and its impact on the livelihood of people living around the lake can be well analyzed and appropriate inferences made from the available data obtained from reliable sources. In this study the problem to be addressed is to estimate the missing values and use the estimates to forecast with a view to bridge the gap whose records were not captured in the time period stated earlier.

1.3 Objectives of the study

1.3.1 General objective

The general objective of the study is to fill the seasonal gap of four consecutive months in the periods stated above.

1.3.2 Specific objectives

- To identify the nature of the data before and after the missing gap,so as to identify a suitable model for the seasonal time series.

- To use the Seasonal Autoregressive Integrated Moving Average SARIMA to estimate the missing values through forecasting.
- To test whether the model so considered provide accurate estimate or otherwise.

1.4 Significance of the study

The study of seasonal variations enable one to understand effects of seasonal patterns on the long term variations about fishing activity in Lake Victoria. Progressive economic development requires more scientific knowledge of relationship between productions at a given time intervals. More accurate predictions would therefore reduce the range of uncertainty and make scientific analysis more valuable and near to reality. The proposed study would therefore provide means to quantify the economic gains foregone in case of calamities that are unpredictable not only in the fishing industry but also in other fields of scientific research. The method used would act as extension to the existing models for better predictions as reliable estimates can also help in economic and social planning for future operations.

1.5 Basic concepts and definitions

1.5.1 Time series

A Time series is some sequence of observations proceeding through time where the actual order in which the values occur has importance in that they may be related temporarily. In this situation the observations are not assumed to be independent. Time series is generally composed of the following components; trend, seasonal effect, cyclic effect and random effect.

1.5.2 Stationary and Stochastic process

A time series 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$(t_i \in \mathbb{R})$.
As a result the parameters such as the mean and variance if they exist do not change over time.It refers to a flat looking series; without trend, with constant variation over time,a constant autocorrelation structure over time and has no periodic fluctuations. A time series is said to be covariance stationary in weak sense if its mean is constant and autocovariance is independent of distance between the variables, i.e $E(X_t) = \mu < \infty$ for all $t \in \mathbb{R}$

$$E(X_t - \mu)(X_{t+h} - \mu) = Cov(X_t, X_{t+h}) = \delta(h) \tag{1.1}$$

Where μ is a real number and $\delta(h)$ is the autocovariance function for lag of h. Note that the autocorrelation function also depends on lag h.Non stationary time series is most commonly used in economic data and its raw data can be transformed to become stationary.The transformation involves differencing the data with the given series X_t to create the new series

$$\bigtriangledown X_t = X_i - X_{i-1} \tag{1.2}$$

The differenced data will then contain less points than the original data. This can be done continuously till a stationary series is obtained.

1.5.3 Autocovariance and Autocorrelation Function

Autocorrealtion function are values that fall between -1 and +1 calculated from time series at different lags to measure the significance of correlations between present and past observations and to determine how far back in time they are correlated.
For a stationary process X_t we have the

mean

$$E(X_t) = \mu \tag{1.3}$$

and variance

$$\delta(0) = E(X_t - \mu)^2 = \delta^2 \tag{1.4}$$

which are constants.Then

$$\delta(h) = Cov(X_t, X_{t+h}) = E(X_t - \mu)(X_{t+h} - \mu) \tag{1.5}$$

is called the autocovariance function which is the function of the time difference $(t, t+h)$. The function

$$\rho(h) = \frac{Cov(X_t, X_{t+h})}{\surd Var(X_t)Var(X_t)} = \frac{\delta(h)}{\surd \delta^2(X_t)\delta^2(X_{t+h})} = \frac{\delta(h)}{\delta(0)}. \tag{1.6}$$

is referred to as Autocorrelation function (ACF) in time series analysis as they represent the covariance and correlation between X_t and X_{t+h} from the same process separated by the time lag h.

Note that for stationary time series the autocovariance and autocorrelation functions have the following properties;

- for uncorrelated data the autocorrelation function is equal to zero i.e $\rho(h) = 0$ for all$h \neq 0$
- $\delta(h) = \delta(-h), \qquad \rho(h) = \rho(-h)$ hence positive half of the autocovariance is commonly plotted in correlogram.
- $-1 \leq \rho(h) \leq 1$

1.5.4 Partial Autocorrelation Function

Partial autocorrelation function values are the coefficients of linear regression of the time series using its lagged values as independent variables. It is equally useful in making time series models especially where there are large portions of correlations between X_t and X_{t+h} in which autocorrelation patterns are difficult to establish. The lag h for partial autocorrelation is the partial regression coefficient Φ_{hh} in the r^{th} order autoregression.

$$X_{t+h} = \Phi_{h1}X_{t+h-1} + \Phi_{h2}X_{t+h-2} + --- + \Phi_{hh}X_t + e_{t+h} \tag{1.7}$$

where e_{t+h} is normal error term.

Multiplying the equation above by X_{t+h-j} and taking the expectation the result is;

$$\delta(j) = \Phi_{h1}\delta(j-1) + \Phi_{h2}\delta(j-2) + \ldots + \Phi_{hh}\delta(j-h) \tag{1.8}$$

Dividing both sides by $\delta(0)$ we get

$$\rho(j) = \Phi_{h1}\rho(j-1) + \Phi_{h2}\rho(j-2) + \ldots + \Phi_{hh}\rho(j-h) \tag{1.9}$$

$j \geq 1$

These correlation functions can be applicable mostly in modelling of statistical dependencies about evolution of time series X_t and therefore can form the basis of rules for interpolating values at points that are lacking data. They play an important role in data analysis aimed at identifying the extent of the lag in autoregressive model. This was introduced by Box-Jenkins (Chapter $3.2, 2008$) as a suitable approach to time series modelling where partial autocorrelation function are plotted to determine the appropriate lags p in an AR(p) models hence in an extended ARIMA models.

An approximate test that a given partial correlation is zero at a(5% level of significance) is given by comparing the sample autocorrelations against the critical region within the limits given by $\pm 1.96/\sqrt{n}$ where n is the recorded number of points of the time series being analysed.

Note that computer software is available to do the production of ACF and PACF.

1.5.5 Correlogram

Is an important tool in time series analysis that can be used to describe the nature of time series and also to identify an appropriate model for a given time series.

The autocorrelation $r_h = \frac{\delta(h)}{\delta(0)}$ where $\delta(h) = \frac{\sum_{t=1}^{N-h}(x_t-\bar{x})(x_{t+h}-\bar{x})}{N}$ for $h = 0, 1, 2 \ldots$ and $\delta(0) = \delta^2$ can be used to determine the correlograms.

When a graph of r_h against h is plotted a correlogram is produced which can assess the behaviour and properties of the time series.

A correlogram can be plotted for the original time series and/or for the differenced stationary series. When time series is stationary the correlograms provide estimates of the theoretical autocorrelation functions but for non-stationary series they do not estimate hence can be used to show that time series is non-stationary.Seasonal series have large values of r_h at the seasonal periods hence one can use the correlograms to observe if seasonality is present.

1.5.6 Moving Average process

Suppose $e(_t)$ is a white noise with mean zero and variance δ^2 then the process X_t is said to be a moving average process of order q if

$$X_t = \beta_0 e_t + \beta_1 e_{t-1} + \ldots + \beta_q e_{t-q} \tag{1.10}$$

where $\beta_0, \beta_1, \beta_q$ are moving average parameters.

The subscripts on the β, s are called the orders of Moving average parameters. The highest order q is referred to as the order of the model, hence can be abbreviated MA(q) which means Moving Average of order q.

Basic model for the moving average is

$$X_t = e_t + \theta_t e_{t-1} \tag{1.11}$$

which indicates that any given current observation is directly proportional to the random error e_{t-1} from preceding period together with the current one e_t. The θ_s refer to the order of the Moving Average parameters.

For MA(q) process

$$\delta(h) = Cov(X_t, X_{t+h}) \tag{1.12}$$

$= E(X_t, X_{t+h}) - E(X_t)E(X_{t+h})$

$= E(X_t, X_{t+h})$ since $E(X_t)E(X_{t+h}) = 0$

But $X_t = \sum_{j=0}^{q} \theta_j e_{t-j}$ and $X_{t+h} = \sum_{j=0}^{q} \theta_j e_{t+h-j} = 0$

$X_t, X_{t+h} = (\sum_{j=0}^{q} \theta_j e_{t-j})(\sum_{j=0}^{q} \theta_j e_{t+h-j})$

$= \sum_{j=0}^{q} \sum_{k=0}^{q} \theta_j \theta_k e_{t-j} e_{t+(h-j)}$

If $j = k - h$,then $e_{t-j} = e_{t-(k-h)}$

$E(e_{t-j}e_{t-(k-h)} = E(e_{t-j})^2 = var(e_t) = \delta^2$

If $j \neq k - h$ then $e_{t-j} \neq e_t - (k-h)$ and $E(e_{t-j}e_{t-(k-h)} = E(e_{t-j})(E(e_{t-(k-h)}) = 0$

Thus

$$\sigma(h) = E(X_t, X_t + h) = Cov(X_t, X_{t+h}) = \delta^2 \sum_{j=0}^{q} \theta_j \theta_{j+h} \quad h = 0, 1, 2 \ldots q \tag{1.13}$$

ACF for an MA(q) is

$$\rho(h) = \begin{cases} 1 & \text{if h=0} \\ \frac{\sum_{j=0}^{q-h} \theta_j \theta_{j+h}}{\sum_{i=0}^{q} \theta_i^2} & \text{h} = 1, 2, \ldots q \\ 0 & \text{h} > q \end{cases}$$

If a correlogram for ACF is plotted then the curve "cuts off" at lag q which is a special feature of MA process.

1.5.7 Autoregressive process

Let (e_t) be a purely random process with mean zero and variance σ^2 , then the process X_t is said to be an autoregressive process of order p if

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + e_t \tag{1.14}$$

where ϕ_1, ϕ_2, ϕ_p are parameters of autoregressive and the subscripts $1, 2, \ldots$ are the orders of the autoregressive parameters which increase with increase in X_t.

The basic model of AR is

$$X_t = \phi X_{t-1} + e_t \tag{1.15}$$

where e_{ts} is a sequence of independent identically distributed normal random variables with mean zero and variance σ^2

It indicates that the value X_t depends directly on previous value of X_{t-1} plus random error e_t.

And as the number of AR parameters increase,X_t becomes directly related to increased past values leading to an expression in (1.15) above and the model looks like a regression model,hence the term autoregression.

The values of ϕ which would make the process to be stationary are such that the roots of $\phi(B) = 0$ lie outside the unit circle in the complex plane. B is the backward shift operator such that $B^j X_t = X_{t-j}$ and $\phi B = 1 - \phi_1 B \ldots \phi_p B^p$ (Chartfield 1989 p.41)

1.5.8 Autoregressive Moving Average

This is a combination of Autoregressive and Moving average models to build a stochastic model that can represent a stationary time series. The order of ARMA are expressed as

p and q respectively and they relate to what happen in period t to the past values and random errors that occurred in the past periods.

The model is

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + e_t - \sum_{j=1}^{q} \theta_j e_{t-j} \tag{1.16}$$

Thus $X_t = \phi_1 + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} - \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_q e_{t-q}$

When simplified using backward shift operator

$B^j X_t = X_{t-j}$

we get

$$\phi_(B)X_t = \theta(B)e_t \tag{1.17}$$

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

$\phi(B) = 1 - \phi_1 B \ldots \phi_p B^p$ are polynomials of degree p and q in B and the process is ARMA(p, q)

To attain stationarity for this model the equation $\phi(B)$ has its roots lying outside the unit circle and $\theta(B) = 0$ must lie outside the unit circle for the process to be invertible.

1.5.9 Autoregressive Integrated Moving Average

If the observed series is non stationary in the trend then we can difference the series to obtain stationarity. In ARIMA models the term integrated, which is acronym for summed is used because the differencing process can be reversed to obtain the original time series values by summing the successive values of the differenced series(Hoff,1983,p.126).

Consider the Autoregressive Moving Average of order (p, q) given by

$$X_t = \phi_1 X_{t-1} + \ldots \phi_p X_{t-p} e_t + \beta_1 e_1 + \ldots + \beta_q e_{t-q} \tag{1.18}$$

Suppose X_t is non-stationary but ∇X_t is stationary

Let $\omega_t = \nabla^{\mathrm{d}} X_t$, ω_t is stationary while X_t is non-stationary. where

$\omega_t = \phi_1\omega_1 + \ldots \phi_p\omega_{t-p} + e_t + \beta_1 e_1 + \ldots + \beta_q e_t - q$. we can write using backward shift operator

$$\Phi(B)\omega_t = \Theta(B)e_t \qquad (1.19)$$

This implies

$\Phi(B)\nabla^d X_t = \Theta(B)e_t$

Note $\nabla = 1 - B$, $\quad \nabla X_t = X_t - X_{t-1}$

$(1 - B)X_t = X_t - (B)X_t$.

Thus

$$\Phi B(1 - B)^d X_t = \Theta(B)e_t. \qquad (1.20)$$

which is autoregressive integrated moving average of order (p, d, q).

1.5.10 Seasonal Autoregressive Integrated Moving Average

Most of the economic time series do show seasonal fluctuations with some characteristics of homogeneity within given periods of the year. The pattern developed can be at intervals of monthly $(s = 1)$,quarterly $(s = 4)$ or yearly$(s = 12)$. When these data are arranged in tabulated form then some relationships can be noted between observations among the same months and also among the successive months of the year. This scenario can be expressed by a model

$$\phi_p(B)\Phi P(B^s)\nabla^d\nabla_s^D X_t = \theta_q(B)\Theta_Q(B^s)e_t \qquad (1.21)$$

where the subscripts $\phi_p\Phi_P\theta_q\Theta_Q$ are polynomials of the corresponding order p, P, q, Q respectively.

And ∇^d is the simple differencing operator of order p and ∇^D is the seasonal differencing operator of order D.

$\nabla^1\nabla = X_t - X_{t_1} = (1 - B)X_t$

Thus the final differenced stationary time series is not only from simple differencing to remove the trend but also seasonal ∇_s to remove seasonality.

Chapter 2

Literature Review

2.1 Introduction

In this chapter related research about time series with missing observations are discussed. It is noted that there have been more revelations on the study of time series with linear observations than non-linear time series.However most of the time series processes are non-stationary which implies that they have mean,variance and autocovariance of the process that are variant under time translations.

2.2 Literature review

In order to remove non-stationary sources of variation and fit stationary models Box and Jenkins in 1976 [5] recommended the extension to Autoregressive Moving Average process which deals with stationary process through differencing.

Richard H. Jones in 1980[19] came up with a method that involved calculating exact likelihood function of stationary Autoregressive Moving Average based on Akaike's Markovian[2] representation combined with Kalman recursive estimation. This approach involved use of matrices and vectors with dimensions equal to max(p,q) where p is the order of autoregressive and q is the order of moving average. His article also mentions some more discussions on observational error in the model and the extension to missing observations.In the same year A.C.Harvey and G.D. A Phillips came up with an algorithm that enables the exact likelihood function of a stationary autoregressive moving average process to be calculated by use of Kalman filter. This involved two procedures; The first one translates autoregressive moving average process model into "state space" form which is necessary for Kalman filtering and the second computes the covariance matrix related with initial values of the state vector.

Priestly in 1981 discussed stationary process as a sum of deterministic and non-deterministic processes. Where deterministic refers to a situation in which the forecast is done by linear regression on past values without necessarily involving recent values. And if the future values are considered to be a realisation from probability distribution which is conditioned by knowledge of past values then the process is non-deterministic(stochastic).

In 1984, A.C.Harvey and R.G. Pierce[8] discussed related problems about time series with missing data. The problems were about use of maximum likelihood estimation of missing observations. They suggested setting up of the model in the state space form and applying Kalman filter.

F.C Ansley and Robert Kohn in the year 1986[3] showed how to define and compute efficiently the marginal likelihood of autoregressive moving average model with missing data using modified Kalman filtering process they developed earlier.They also showed how to predict and interpolate missing observations and to obtain mean squared error of the estimate.

In 1989 Greta M. Lyung[25] came up with the expression for the likelihood function of parameters in Autoregressive intergrated Moving Average model when there are missing values within time series data.

In 1991 Daniel Rena and George C Tiao demonstrated that missing values in the time series can be treated as unknown parameters and estimated by maximum likelihood or as random variables and predicted by the expectation of the unknown values given the data.

In 1996 Fabio H. Nieto and J. Martinez demonstrated a linear recursive technique that could be used to estimate missing observations in a univariate time series without use of Kalman filter. It focusses on forecasting approach and the recursive linear estimators obtained when the minimum mean square error are optimal.

In 1997 Albert Luceno extended Lyung's method of estimating the corresponding likelihood function in scalar time series to the vector cases. Here the series assume no particular pattern of missing data existed. It does not require the series to be differenced hence avoiding the complications that could arise by over differencing. The estimators of the missing data are provided by the normal equations of an appropriate regression technique.

In the year 2003 Chris Chatfield[10] in his article entitled"Analysis of time series an introduction" gave various approaches for linear time series most of which involved curve fittings. From the above literature it is noted that a lot of methods have been developed that could be applied to address the problem of estimating missing values in a time series.However our problem would require use of the most relevant approach that could be used to convert non-stationary time series to stationary model by use of seasonal autoregressive moving average.

2.3 Research Methodology

The preliminary stages of this study focuses on the use of Box-Jenkins Autoregressive Integrated Moving Average model to identify the most suitable model for the data obtained.This is because Box-Jenkin's models are able to handle varieties of non-stationary time series by differencing to attain stationarity.It can effectively deal with time series that have historical behaviour, which is common with economic time series such as fish harvesting.It develops the model in a systematic form that is easy to follow and the model so developed can be systematically tested.This study focuses on the seasonal data analysis. It will involve observing the nature of the time series from the plotted graph. It will be tested for stationarity and linearity using a suitable test. Once confirmed the appropriate model that can fit the data would be identified.Parameters of the identified model is to be determined and a diagnostic model checking be done to ensure that it can be adequately applied in the obtained time series data.

Using the model forecast will be done on the lower segment of the observed values and the level of accuracy noted before the forecast is applied to the actual missing values i.e between the months of December 2007 and March2008. forecasting would be prolonged past the gap and through regression analysis the level of accuracy be noted. With test of accuracy done conclusion to be made will highlight on the performance of the model and its applicability in forecasting to periods outside the observed data.It is hoped that the results obtained and the consequent suggestions would be useful to both short term and long term research activities whose outcome would have greater certainty.

Chapter 3

Modelling of the time series

3.1 Introduction

A model is a mathematical expression representing reality. The model can be used to describe the time series data. It can be used to determine theoretical relationships between different sets of data. And it can also be used to predict the unobserved variations.The process of analysing and forecasting time series requires that a model for a suitable set of data is identified and used. Many models can be produced from a given set of data. This chapter therefore tries to address this important aspect of analysis by identifying the most appropriate model that suits the data set and use it to predict missing values. Fitting a model of time series takes place in stages.
These include;

- Model identification
- Estimation of the model parameters
- Diagnostic checking of the model.

However before identification is done it is important to show the nature of the data i.e whether the distribution pattern of the data shows linearity or non linearity.After this then the model that can fit the pattern is possible to identify.

3.2 Linear and non-linear time series model

Generally a linear time series model takes the expression of the form

$$\sum_{i=0}^{\infty} h_i X_{t-i} = e_t \tag{3.1}$$

Where e_t is the white noise.

Using backward shift operator $B^j X_t = X_{t-i}$

then equation above becomes

$H(B)X_t = e_t \ldots$

Thus

$$X_t = H^{-1}(B)e_t = (m_0 + m_1 B + m_2 B^2 + \ldots)e_t \tag{3.2}$$

$$X_t = \sum_{i=1}^{\infty} m_i e_{t-i} \tag{3.3}$$

The above equation gives a general linear model where by the X_t is expressed in terms of the present and past values of the white noise process. These types of models can be fitted for AR, MA and ARMA.

For non-linear models, according to Priestly(1980,1988) is of the general form

$$X_t = \mu + \sum_{i=0}^{\infty} m_i e_{t-i} + \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} m_{ij} e_{t-i} e_{t-j} + \sum_{i}^{\infty}\sum_{j}^{\infty}\sum_{k}^{\infty} m_{ijk} e_{t-i} e_{t-j} e_{t-k} + \ldots \tag{3.4}$$

where $\mu = f(0)$ $\quad m = \frac{\Delta f}{\Delta e_{t-i}}$ where $X_t = f(e_t, e_{t-1}, e_{t-3} \ldots)$

From the above equation non-linearity exists if the higher order coefficients $(m_{ij})(m_{ijk})$ are non zero.Our data forms linear time series that is non stationary.

3.3 Linear non-stationary time series model

Many economic activities do form time series patterns which,although are non-stationary,do exhibit some homogeneity with predictable repeated patterns which if differenced to an appropriate degree can be converted to stationarity.The key requirement is that time series is either stationary or can be transformed to into one. A plot of the data is usually enough to verify if the data is stationary or not. However in practice few time series meet this condition hence the need to transform the data into stationary series.

The models for the stationary series are of the general form

$$\phi(B)(1-B)^d X_t = \theta(B)e_t \tag{3.5}$$

Where for time interval T,$X_t(t\epsilon T)$ is a sequence of random variables,B is the backward shift differencing operator.

$B^h X_t = X_{t-h}$ (h is non-negative integer),

$(1-B)^d X_t = \nabla^d X_t$ is stationary,

$\phi(B)$and $\theta(B)$ are linear filters defined as

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

$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q$

and $e_t(t\epsilon T)$ is a sequence of uncorrelated random variables with mean zero and variance σ^2 (white noise).

In case the series show some seasonal fluctuations within the year together with trend then the above model is modified appropriately. Box-Jenkins came up with a general model that covers seasonality as given

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D X_t = \theta(B)\Theta(B^s)e_t \tag{3.6}$$

referred to as Seasonal Autoregressive Integrated Moving Average(SARIMA),where X_t, e_t, $\phi(B)$ar $\theta(B)$ are defined as above.
$(1-B)^d(1-B^s)^D X_t = \nabla_1^d \nabla_s^D X_t$ is stationary and $\Phi(B^s)$and$\Theta(B^s)$ are seasonal linear filters defined as
$\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \ldots - \Phi_p B^{ps}$ and
$\Theta(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \ldots + \Theta_Q B^{QS}$.
P defines the seasonal autoregressive component of the model and Q defines the seasonal moving average component of the model and s represents the seasonal period and D is the degree of seasonal differencing. Note that both S and D account for seasonal non-stationarity in X_t.

It is noted that SARIMA model is an extension of ARIMA model and its modelling follows the same procedures of identification,estimation and checking.

3.4 Identification of the model(p,d,q)

This is a procedure meant to select a model that is more suitable for the data obtained. It should be noted however that this stage may not give the exact model but can provide a class of model which could be verified further.
Use of correlogram of the data indicates whether the series is stationary or non-stationary as non-stationary data has the correlogram that fails to decay to zero.
A correlogram of a time series$(X_t : t = 1, 2, \ldots n)$ is a graph of the sample autocorrelation coefficient τ_h against the corresponding lags h.Each τ_h is defined as

$$\tau_h = \frac{g_h}{g_0} = \frac{\sum_{t=h+1}^{n}(x_t - \overline{x})(x_{t-h} - \overline{x})}{\sum_{t=1}^{n}(x_t - \overline{x})(x_t - \overline{x})} \tag{3.7}$$

The level of individual departure of τ_h is checked within the limits of $\pm 2/\sqrt{n}$. Also the pattern formed by τ_h can be used to determine the nature of time series by examining the points at which the τ_h " cuts off " the point zero within the limits.

For all τ_h where $h > q = 0$ indicates that MA(q) is a suitable model and for AR process the autocorrelations decaying exponentially is observed by partial autocorrelation function which has a cut off for an underlying process at $\phi_{hh} = 0 \quad \forall h > p$

According to Chartfield (1975) the seasonal time series should be differenced (d times to remove trend and D times to remove seasonality) so as to reduce it to stationarity. The general SARIMA model is of the form

$$\phi_p(B)\Theta_P(B^s)w_t = \theta_q(B)\Theta(B^s)e_t \tag{3.8}$$

w_t is the result of differencing X_t till when its autocorrelation function dies out quickly. d usually takes values $0, 1, 2$. The values p, P, q, Q are determined from the patterns from SACF and SPACF of the differenced series. P and Q are examined from $\tau_h = s, 2s$ where s is the periods. This identified model can then be compared with theoretical patterns of known models as shown in the appendix.
In summary the data is AR(p) if its ACF will decline steadily, or follow a damped cycle and PACF will cut suddenly after p lags. It is a MA(q) if its ACF will cut off suddenly after q lags and PACF will decline steadily or follow a damped cycle.

It should be noted that model identification by Box-Jenkins method is considered subjective due to the fact that it primarily relies on graphical interpretation of ACF/PACF estimates from a single sample. The minimum sample size generally recommended for the SARIMA model fitting is 50 observations(Pancratz 1983;Chatfield 1996). And as the sample size become larger ACF/PACF estimates tend to lower variability hence better approximation of the underlying process. However when the sample size is small then the interpretation of ACF/PACF patterns will acquire larger variances leading to subjectivity of the model identification.
To reduce this subjectivity, a model selection criteria referred to as Akaike Information Criterion(Akaike,1974)and the small sample bias corrected equivalent AIC_c(Hurvich and Tsai,1989)is used. Bayesian Information Criterion (BIC)can as well be used.

AIC/AIC_c selection of the model involves estimation by maximum likelihood methods of a set of model candidates. The model candidates will then have their AIC/AIC_c values determined and the model candidates with minimum AIC/AIC_c is then selected as the model that is closest to the statistical process generating the data.

AIC is calculated as $AIC = -2ln(L) + 2r$ where ln(L)is the loglikelihood of the model and $r = p + q + P + Q + 1$

$AIC_c = -2ln(L) + 2r + 2r(r + 1)/(n - r - 1)$ where $n = N - D - d$ is the number of observations used to fit the model.

And BIC$= -2ln(L) + r + rlnN$

Both AIC and BIC involves objective approach with adequate penalty terms to models with excessive model parameters. It thus encourages a model with fewer parameters.

3.5 Model estimation

After identifying the model its parameters are estimated. The following discussion tries to estimate the model parameters for ARMA model.

Consider an AR process of order 1 given as

$$(X_t - \mu) = \alpha(X_{t-1} - \mu) + e_t \tag{3.9}$$

We wish to estimate μ and α from the observed series.

We can give maximum likelihood approach so as to estimate the above parameters.

We note that from the above equation:

- μ represents the mean value of each X_t hence we estimate $\widehat{\mu} = \overline{x}$ the sample mean of the data.
- α represents the first autocorrelation of (X_t). So we estimate it by $\widehat{\alpha} = \tau_1$, the first sample autocorrelation coefficient.

- Given $\widehat{\mu}$ and $\widehat{\alpha}$ we can construct residuals $e_t = (\widehat{X}_t - \mu) - \widehat{\alpha}(X_{t-1} - \widehat{\mu})$ $\quad t = 2, 3 \ldots n$ and the estimate $\sigma^2 = var(e_t)$ by the residual mean square $\frac{\sum_{t=2}^{n} e_t^2}{n-1}$ S $(\mu, \alpha) + \sum_{t=2}^{n}(X_t - \mu) - \alpha(X_{t-1} - \mu)^2$
  to obtain least μ and α squares we differentiate S (μ, α)

$\frac{\delta S}{\delta \alpha} = 2\sum_{t=2}^{n}(X_t - \mu) - \alpha(X_{t-1} - \mu)(\alpha - 1) = 2(\alpha - 1)[\sum_{t=2}^{n}(X_t - \alpha X_{t-1}) + (\alpha - 1)(n-1)\mu]$

and

$\frac{\delta S}{\delta \alpha}(\mu, \alpha) = -2\sum_{t=2}^{n}(X_t - \mu) - \alpha(X_{t_1} - \mu)(X_{t-1} - \mu)$

$= -2[\sum_{t-2}^{n}(X_t - \mu)(X_{t-1} - \mu) - \alpha\sum(X_{t-1} - \mu)^2]$

for least squares

let $\frac{\delta S}{\delta \mu} = 0 \qquad \frac{\delta S}{\delta \alpha} = 0$

$$\widehat{\mu} = \frac{\sum_{t=2}^{n}(X_t - \widehat{\alpha} X_{t-1})}{1 - \widehat{\alpha})(n-1)} = \frac{-\widehat{\alpha} X_1 + (1 - \widehat{\alpha})\sum_{t=2}^{n-1} X_t + X_n}{(1 - \widehat{\alpha})(n-1)} \tag{3.10}$$

$$\widehat{\alpha} = \frac{\sum_{t=2}^{n}(X_t - \widehat{\mu})(X_{t-1} - \widehat{\mu})}{\sum_{t=2}^{n}(X_{t-1} - \widehat{\mu})^2} \tag{3.11}$$

$$\widehat{\delta}^2 = \frac{\sum_{t=2}^{n}(X_t - \widehat{\mu}) - \widehat{\alpha}(X_{t-1} - \widehat{\mu})^2}{(n-3)} \tag{3.12}$$

For moving average $\hat{\beta}$ is approximated by the solution of the following recursive equations in the form:

$$e_t = X_t - \mu - \beta e_{t-1} \tag{3.13}$$

$$e_1 = X_1 - \hat{\mu} \tag{3.14}$$

$$e_2 = X_2 - \hat{\mu} - \hat{\beta} e_1 \tag{3.15}$$

$$e_n = X_n - \hat{\mu} - \hat{\beta} e_{n-1} \tag{3.16}$$

Then$\sum_{t=1}^{n} e_t^2$ may be calculated for the initial values of $\hat{\mu}$ and $\hat{\beta}$ a procedure that could be repeated for other values of μ and β and the sum of squares $\sum e_t^2$ computed for a grid of points in the (μ, β) plane.

The parameters can be estimated using statistical package.

3.6 Model checking

Before fitting the model to the data it must be checked and modified accordingly to ensure that it is consistent with background knowledge and characteristics of the given data. With residual analysis done

$$e_t = x_t - \widehat{x}_{t-1} \tag{3.17}$$

a good model provides the random series and its autocorrelation function if examined fall within the bounds of $(2/\sqrt{N})$ in absolute magnitude.

Note that inadequate model will fail to provide optimal forecast which is essential for an out of sample predictions. It is therefore important to check the fit of the data by calculating the residuals and then plotting them against time as well as calculating their correlograms.

For serially correlated residuals in ARMA(p,q) Ljung-Box (1978) proposed use of the statistic

$$Q = n(n+2)\sum_{h=1}^{m}(n-h)^{-1}\widehat{(rh)} \sim \chi^2(m-p-q) \tag{3.18}$$

where n is the number of terms in the differenced series, m is the sample size which is usually chosen in the range of 15 to 30. For SARIMA (p,d,q)(P,D,Q) the above statistic remains valid with little adjustment χ^2 having m-p-q-P-Q degrees of freedom. Thus,

$$Q = n(n+2)\sum_{h=1}^{m}(n-h)\widehat{(rh)} \sim \chi^2(m-p-q-P-Q) \tag{3.19}$$

If $Q_{computed} > Q_{table}$ then reject the fitted model.

3.7 Forecasting

3.7.1 Introduction

Here we discuss how the current observed time series can be used to predict future values. This can be done by following the forecasting procedures. The process can be achieved by interpolation of the missing gaps falling between the two sections of the known data. The challenges to this process are:

- whether the historical data can continue to the future predictions
- whether the most appropriate model that can sustain the future predictions accurately
- whether the level of errors that are carried forward may increase in future predictions thereby losing the vital accuracy required.

According to P.Diggle (1990, p.194), Box and Jenkins (1970) built a general forecasting methodology from the assumption that the time series in question possibly after transformation and differencing is generated by stationary autoregressive moving average process. Forecasting procedure therefore involves finding a suitable model, fitting and checking it to examine trend and seasonality. This can be done by use of correlogram and sample autocorrelation function to determine the appropriate model that describes the data adequately.

3.7.2 Imputation method

This is method of filling the missing gaps by putting reasonable values whose characteristics are similar to the observed values. Imputation can be direct or indirect and for ARIMA models,indirect regression imputation method is commonly used. It involves replacing missing observations with predicted values from a model through extrapolation. Abraham(1981) suggested interpolation of adjusted missing value based on the known segments of ARIMA (p,d,q). This can be approached in two ways.

- *Eventual forecast from a model*

  This involves fitting the model to each segment of the known data and extrapolating the missing values from eventual forecast. In this case forecast would be done to the lower segment and back cast to the upper segment. The two estimates if combined should give more accurate estimate.

- *Holt-winter forecasting procedure*

  This procedure is most suitable for linear seasonal time series. In this procedure future observations are made based on the weight of the most recent observations known.

  Thus current observation is estimated as

$$X_n = \sum_{t=0}^{n-1} \omega^t X_{n-t} \tag{3.20}$$

where (ω^t) is a set of weights which sum up to unity.

According to Abraham and Ledolter (1983) k step ahead forecast made at n can be calculated from

$$X_{(n,k)} = C \sum_{t=0}^{n-1} \omega^t X_{n-1} \tag{3.21}$$

where ω is a discount coefficient such that $\omega < 1$ and $c = \frac{1-\omega}{1-\omega^n}$ is a factor needed to normalise the sum of weights to 1

Since

$$\sum_{t=0}^{n-1} \omega^t = \frac{1-\omega^n}{1-\omega} \tag{3.22}$$

it follows that

$$C \sum_{t=0}^{n-1} \omega^t = 1 \tag{3.23}$$

If n is large then the term ω^n in C goes to zero and exponentially weighted 1-step ahead forecast can be written

$$X_{n,1} = (1-\omega) \sum_{j=0} \omega^j X_{n-j} \tag{3.24}$$

The coefficient $\alpha = 1 - \omega$ is referred to as smoothing constant and is usually chosen between 0.05 to 0.3(See Abraham and Ledolter(1983)section 3.3)The above equation can be written as

$$X_{n,1} = \alpha(X_n + (1-\alpha)X_{n-1} + (1-\alpha)^2 X_{n-2} + \ldots) \tag{3.25}$$

$\alpha \sum_{j=1}^{\infty}(1-\alpha)^j X_{n-j}$

$$\alpha(1 + (1-\alpha)B + (1-\alpha)^2 B^2 + \ldots (1-\alpha)^j B^j + \ldots)X_n \tag{3.26}$$

But $(1 + (1-\alpha)B + (1-\alpha)^2 B^2 + \ldots) = (1-(1-\alpha)B)^{-1}$
So that equation above becomes

$$\widehat{X}_{(n,1)} = (\frac{1}{[1-(1-\alpha)B]})\alpha X_n \tag{3.27}$$

This implies $[1-(1-\alpha)B]\widehat{X}_{n-1} = \alpha X_n$

$$\widehat{X}_{(n,1)} = \alpha X_n + (1-\alpha)\widehat{X}_{(n-1,1)} \tag{3.28}$$

From the equation above $\widehat{X}_{(n,1)}$ is weighted average of current observation X_n and the previous forecast $\widehat{X}_{(n-1,1)}$ We can rewrite it

$$\widehat{X}_{(n-1,1)} = \alpha[X_n - \widehat{X}_{(n-1,1)}] \tag{3.29}$$

an equation which can be used to update the smoothed statistics any time t.
This algorithm due to C. Holt is referred to as exponential smoothing.

Chapter 4

Data analysis and Discussion

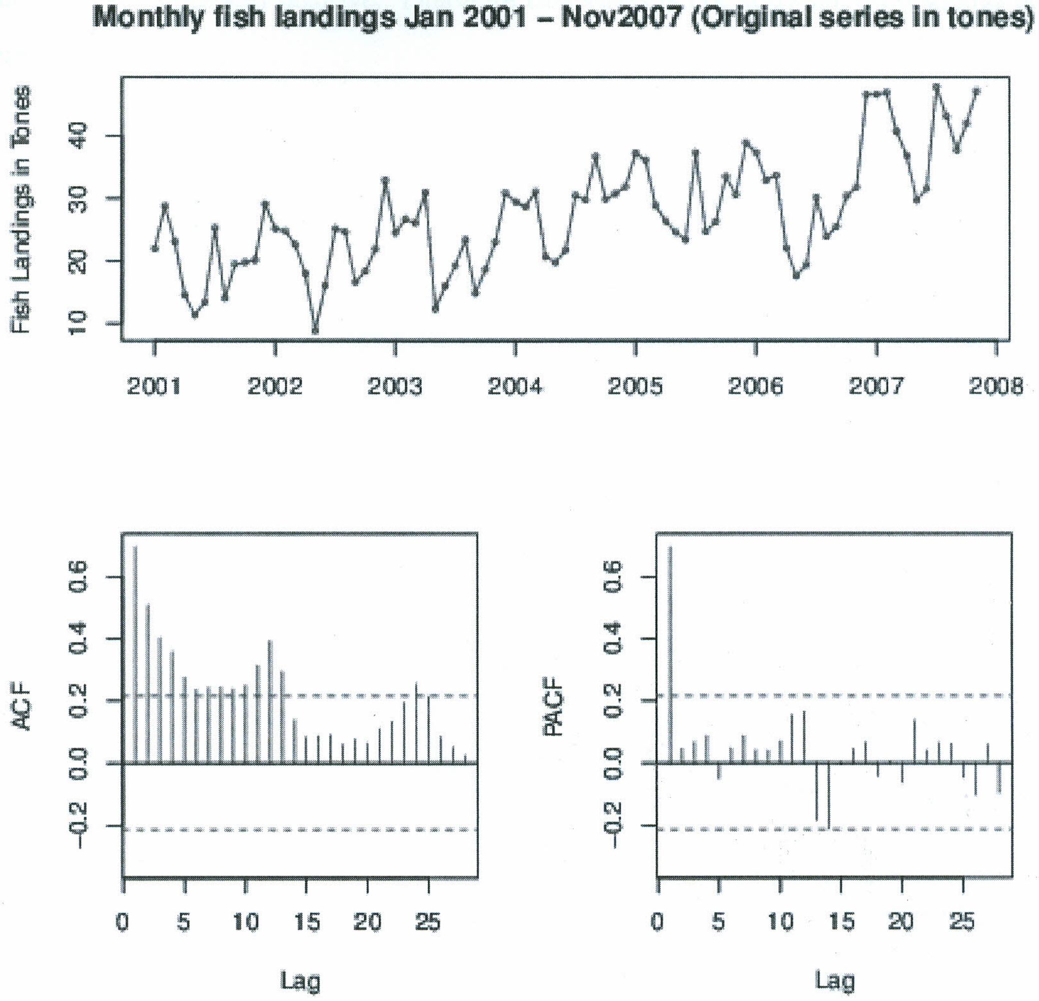
4.1 Introduction

In this chapter analysis is done based on what was discussed in the preceding chapter. This chapter intends to identify the model for both the lower and the upper segment of the time series data.Both would be useful in imputation by forecasting to estimate the four missing observations.

4.2 Data analysis on segment1

Here we examine the plots of the entire data from January 2001 to December 2012 including the missing values. We also examine their ACF and PACF which is useful in determining stationarity/non-stationarity.the results would provide good base to convert it to stationarity for the lower segment with a view to obtain the appropriate model.

Figure 4.2: Time plot between Jan2001-Nov2007(with ACF and PACF).

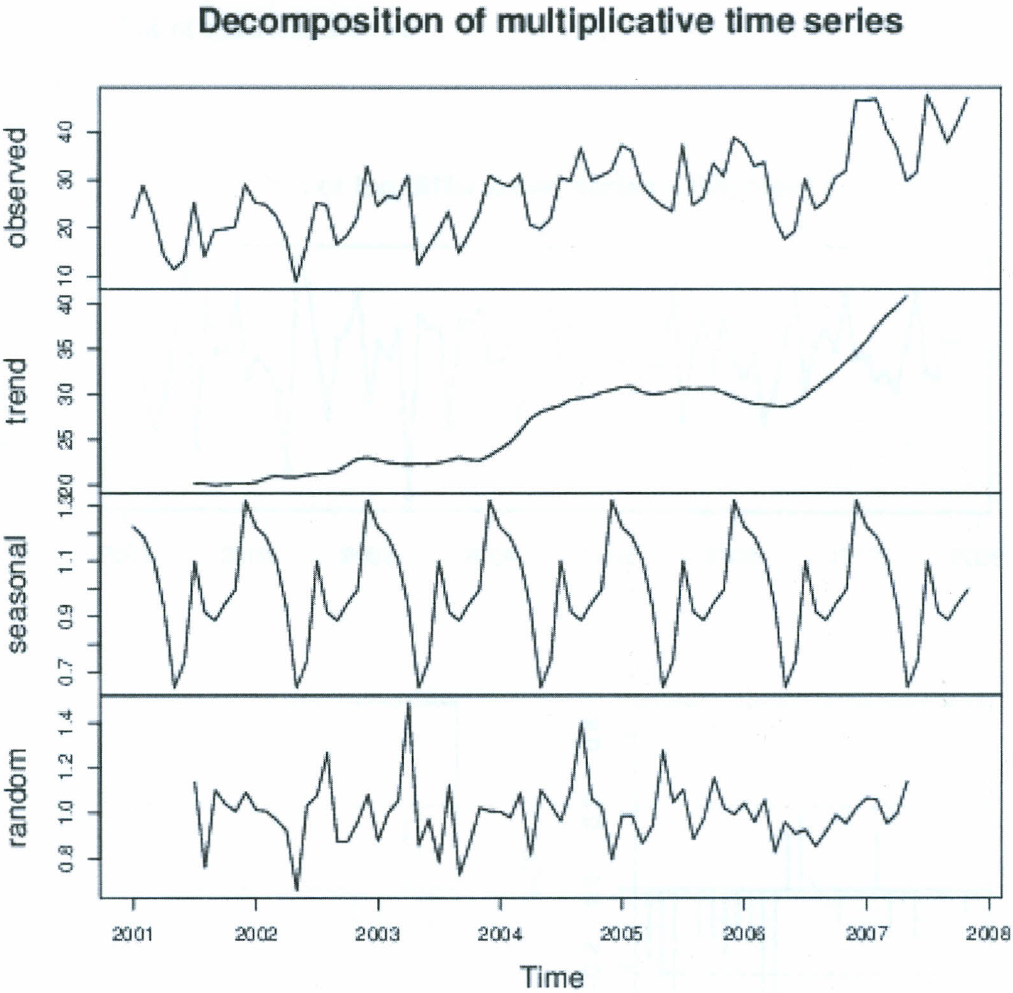


4.2.1 Stationarity

The above time plot for the data for the period January 2001 to November 2007 which is a subset of the entire data set for January 2001 to December 2012. The window function in statistical package R was used to get this subset of the data. There seems to be an increasing trend and seasonal variations in the time series data as shown by the above time plot,plotted along with an ACF and PACF.

From these plots, we see that the fish landing data are seasonal and trending upwards. This means that the mean of the data will change over time.The ACF decreases slowly and they are large and positive. Therefore this series is not stationaryand should be differenced.

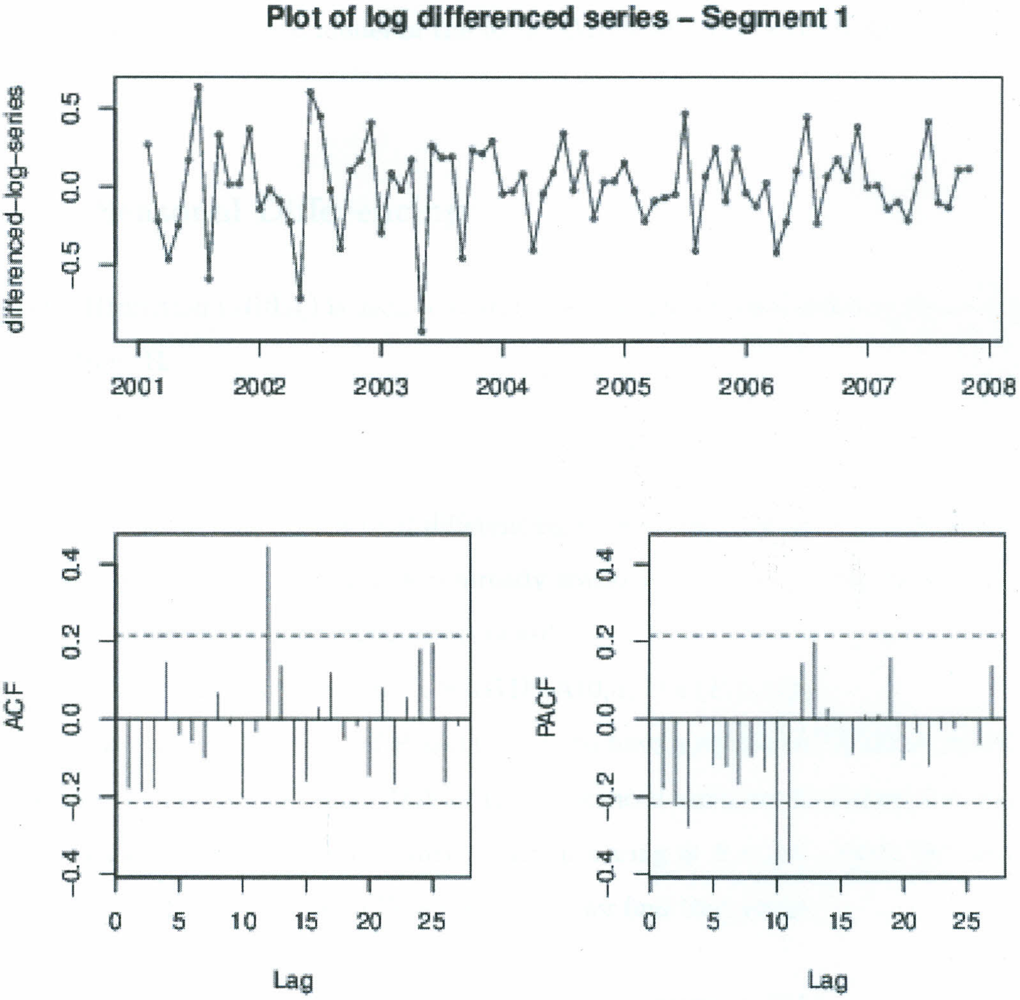
Figure 4.3: Decomposed series of data for (Jan2001-Nov2007)



The aim of decomposition is to separate(estimate) the time series into its three components: seasonal component,trend and irregular(random)component.

We use the multiplicative model since the seasonal and random fluctuations do not seem to be roughly constant over time. To convert it to additive series, we take logs first. The decomposed series is as shown above.The decomposed series shows an overall increasing trend and presence of seasonal variations.

Figure 4.4: Plot of differenced series of segment1(to remove trend component) with ACF and PACF



Stationary test takes the null hypothesis that the time series is trend stationary.

4.2.2 SARIMA Model

The first difference has achieved stationarity in the series. The ACF shows that the differenced series is stationary. This means that the order of non-seasonal differencing shall be 1(d=1). Consequently, we shall have that d=1 in the model to be proposed. This means that the seasonal ARIMA model shall be of the form ARIMA$(p, 1, q) * (P, D, Q)_s$. Where S is the number of observations per period. In this case it is 12 since we have 12 observations per year. Thus we have zero order autoregressive component of the model,AR(0). Therefore so far we have a model of the form SARIMA$(0, 1, 1) * (P, D, Q)_{12}$

4.2.3 Seasonal Differencing

A useful **R**function **ndiffs()** is used to determine whether seasonal differencing is required. Results from **R**:

ndiffs(fishdata1$_{ts}$)

[1]0

This result means that no seasonal differencing is required. This is also seen in the time plot of the differenced series, which is already stationary. Consequently **D** shall be equal to zero(0) in the model that shall be proposed.

The model shall then be of the form **SARIMA**$(0, 1, 1) * (P, 0, Q)_{12}$.

Next, we determine the values of P and Q so as to have a complete SARIMA model. The characteristics of the ACF and PACF of the differenced series tend to show a strong peak at $h = 1s, 2s$ in the ACF, with smaller peaks appearing at $h = 2s$, . Upon the inspection of the ACF and PACF of the differenced series, we find that either:

- The PACF tail off in the seasonal lags. This suggests an SMA of order $Q = 1$
- The PACF has two spikes. This suggests an SAR of order 2 i.e $p = 2$

We therefore have two candidate models:

- (i)SARIMA $(0, 1, 1) * (0, 0, 1)_{12}$
- (ii)SARIMA$(0, 1, 1) * (2, 0, 0)_{12}$

Model[1]:SARIMA$(0, 1, 1) * (0, 0, 1)_{12}$

fishdata1$_{ts}$ $- sarima1$

Series:fishdata1 $- ts$

ARIMA$(0, 1, 1) * (0, 0, 1)[12]$

ma1 coefficient$= -0.4602$ and s.e $= 0.1410$

sma1 coefficient$= 0.4584$ and s.e $= 0.0966$

σ^2 estimated as 31.87 Log likelihood$= -259.81$

AIC$= 525.63$ BIC $= 532.85$

Model[2]:SARIMA$(0, 1, 1) * (2, 0, 0)_{12}$

fishdata1$_{ts}$-sarima2

Series:fishdata1$_{ts}$

ARIMA$(0, 1, 1) * (2, 0, 0)_{12}$

ma1 coefficient$= -0.5620$ and s.e$= 0.1076$

sar1 coefficient$= 0.4914$ and s.e $= 0.1098$

sar2 coefficient$= 0.2388$ and s.e$= 0.1233$

σ^2 estimated as 25.08 loglikelihood$= -252.58$

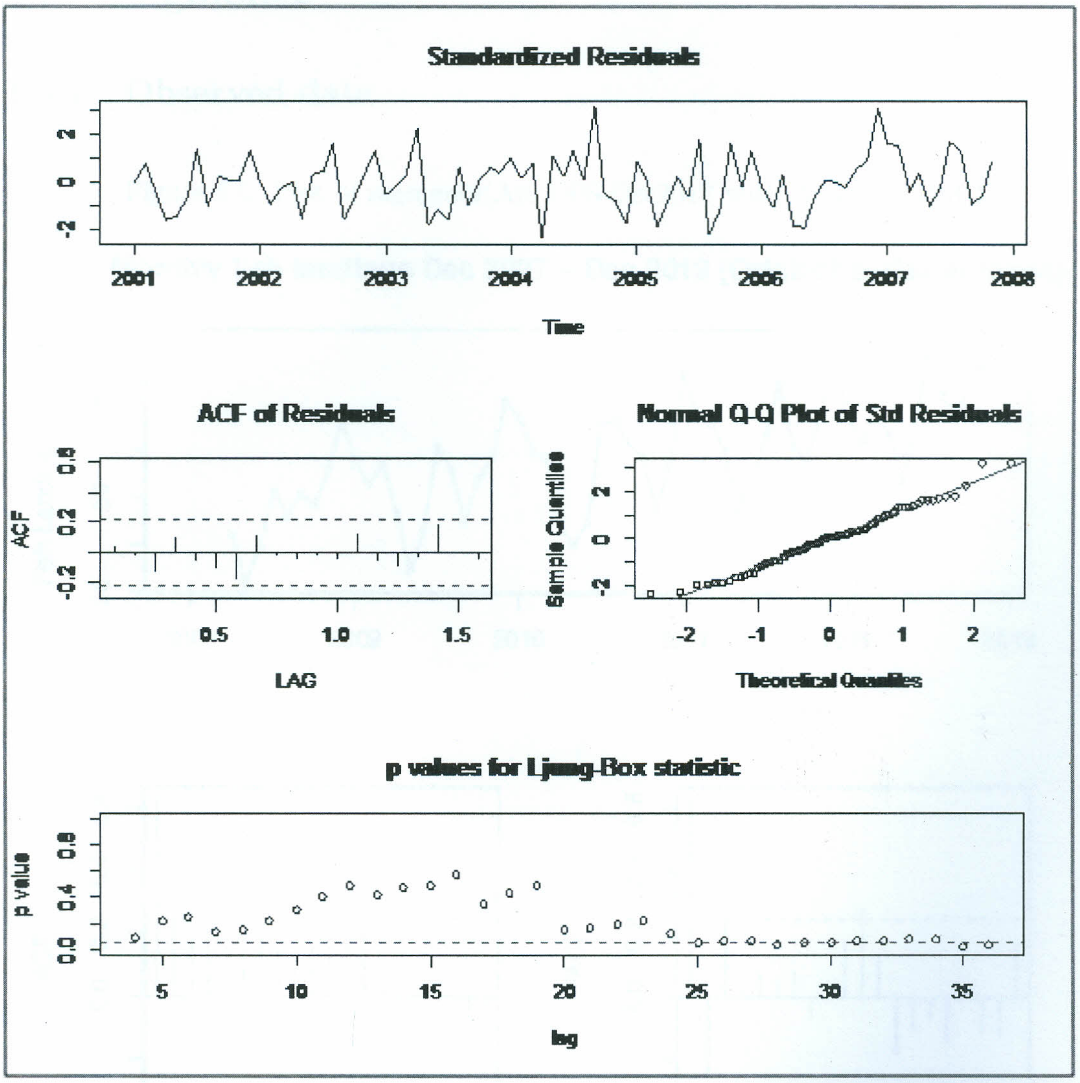
AIC$= 513.16$ BIC $= 522.79$

4.2.4 Decision

Based on the results shown above, we entertain the second model; $\mathbf{SARIMA}(0,1,1)*(2,0,0)_{12}$ since it has smaller AIC and BIC values compared to the first $\mathbf{SARIMA}(0,1,1)*(0,0,1)_{12}$

4.2.5 Diagnostic test

Figure 4.5: Diagnostic test for model adequacy of segment1



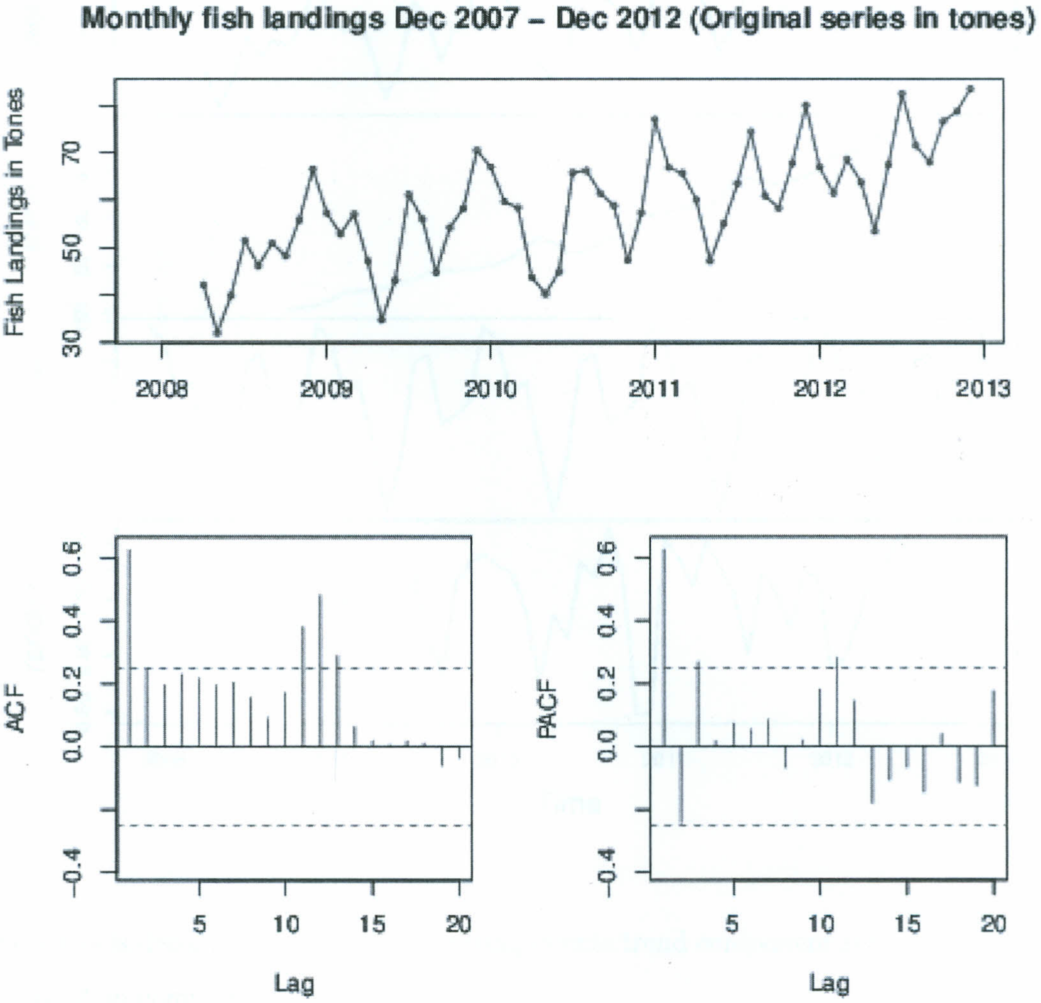
The statistical significance of the derived model must be checked for adequacy. This is done by considering the properties of the residuals from SARIMA model to be normally and randomly distributed. The diagnostic test shows that the model is good. The visual assessment indicate that ACF plot shows none of the sample autocorrelations for lags $1 - 15$ exceed the significance bounds and the Q-Q plot also shows normality.

We can conclude that there is very little evidence for non-zero autocorrelations in the model residuals at lags $1 - 15$. We can therefore proceed to forecasting using the model.

4.3 Segment2

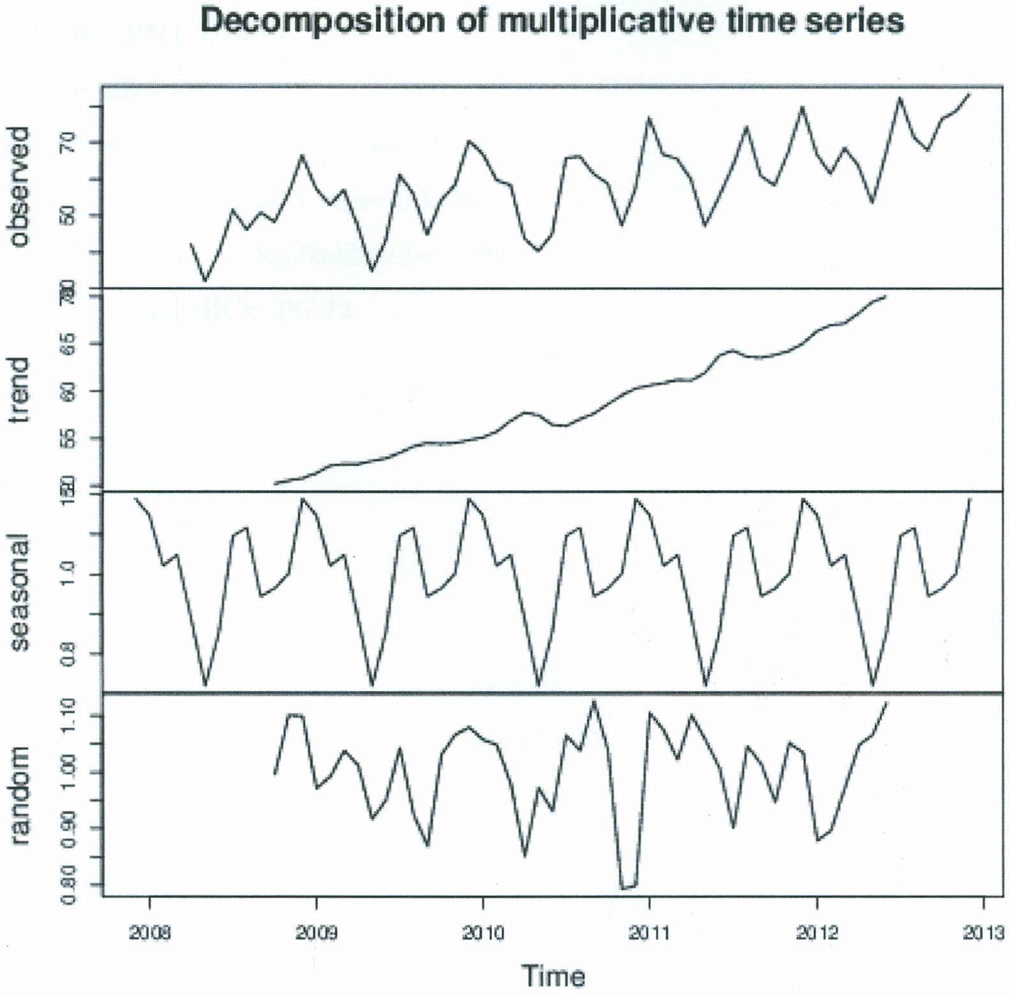
4.3.1 Observed data

Figure 4.6: Plot of segment2(Apr2008-Dec2012)with ACF and PACF



In the above time series data we use auto.arima in **R** to get a suitable SARIMA model for this series.

Figure 4.7: Decomposed data into its components(Apr2008-Dec2012)



The series is decomposed into its three components:trend component,seasonal component and random component.

4.3.2 SARIMA Model

The suitable model identified by the automatic procedure was ARIMA$(0, 1, 2) * (1, 0, 0)_{12}$. The results from **R** are as follows:

auto.arima(fishdata2_{ts})

Series: fishdata2_{ts}

ARIMA(0,1,2)*(1,0,0)-[12]

ma1 coefficients= -0.3314 s.e= 0.1360

ma2 coefficients= -0.4915 s.e= 0.1203

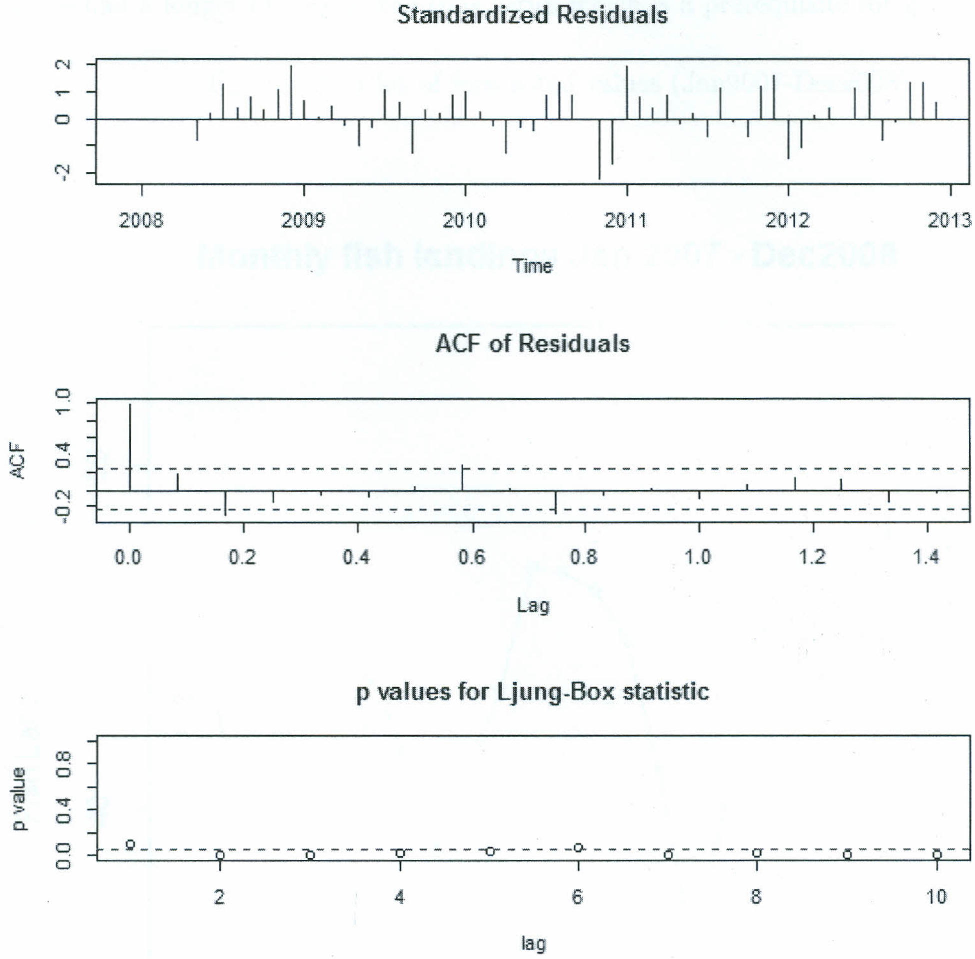
sa1 coefficients= 0.4978 s.e= 0.1324

σ^2 estimated 48.65 loglikelihood= -190.37

AIC= 388.74 and BIC= 397.12

4.3.3 diagnostic test

Figure 4.8: Diagnostic test for model adequacy(segment2)



The model seems to suit the above time series data very well as the p-values are almost near the zero autocorrelations hence within significance bounds.

4.4 Forecast

4.4.1 forecasted values

Values forecasted from the lower segment of the time series data include:54.1, 53.7, 52.8 and 49.9. The lower segment was used to forecast since it had a lot of observed data hence had a longer history of the time series which is a prerequisite for good forecast.

Figure 4.9: Plot of forecasted values (Jan2007-Dec2008)

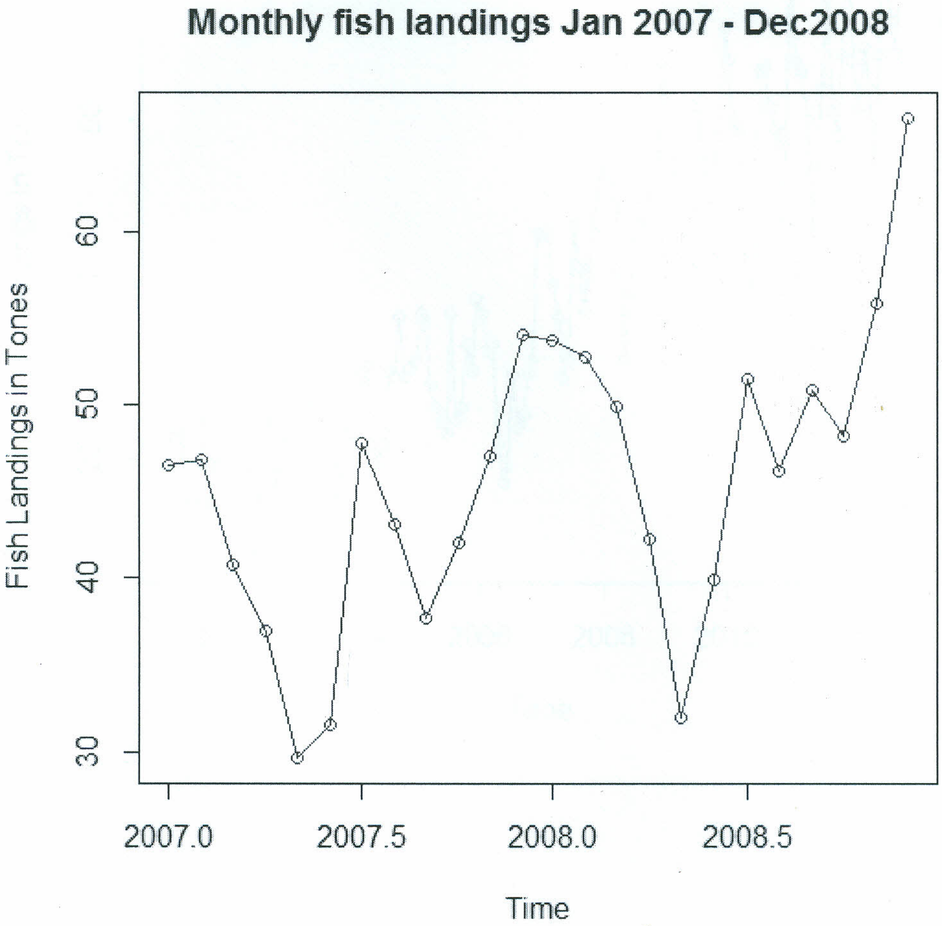


Figure 4.10: Plot of complete fish landings including forecasted values (Jan2001-Dec2012)

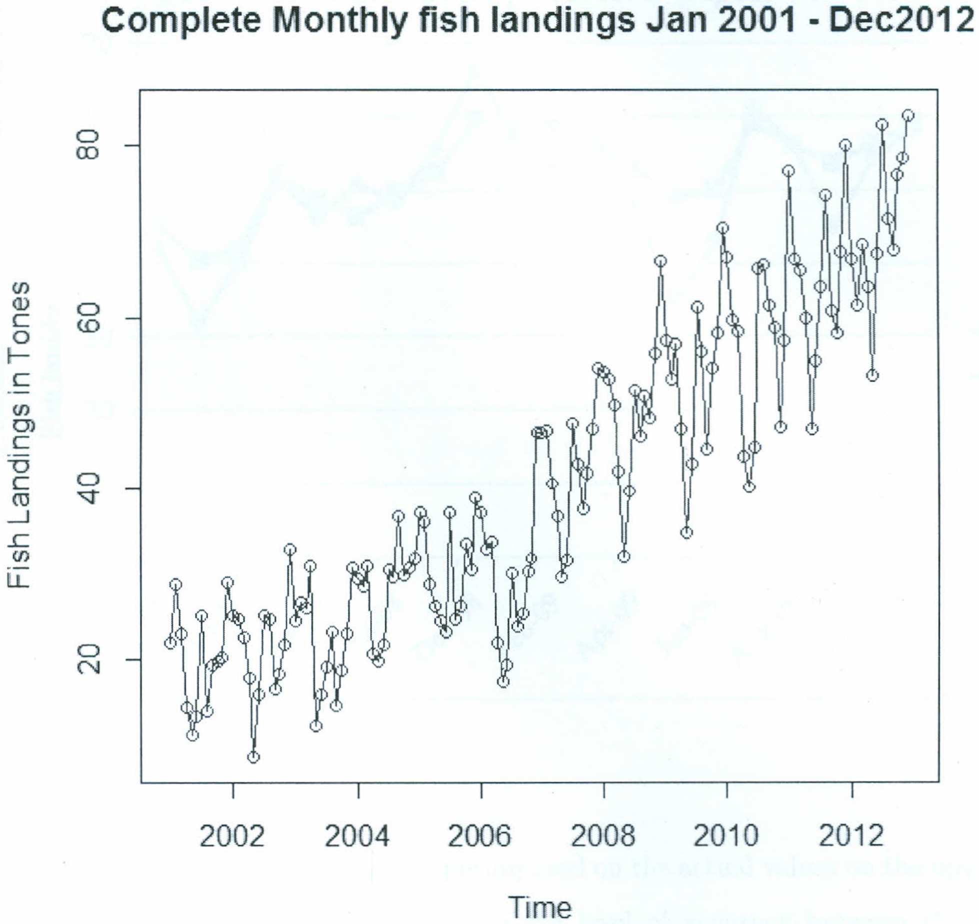
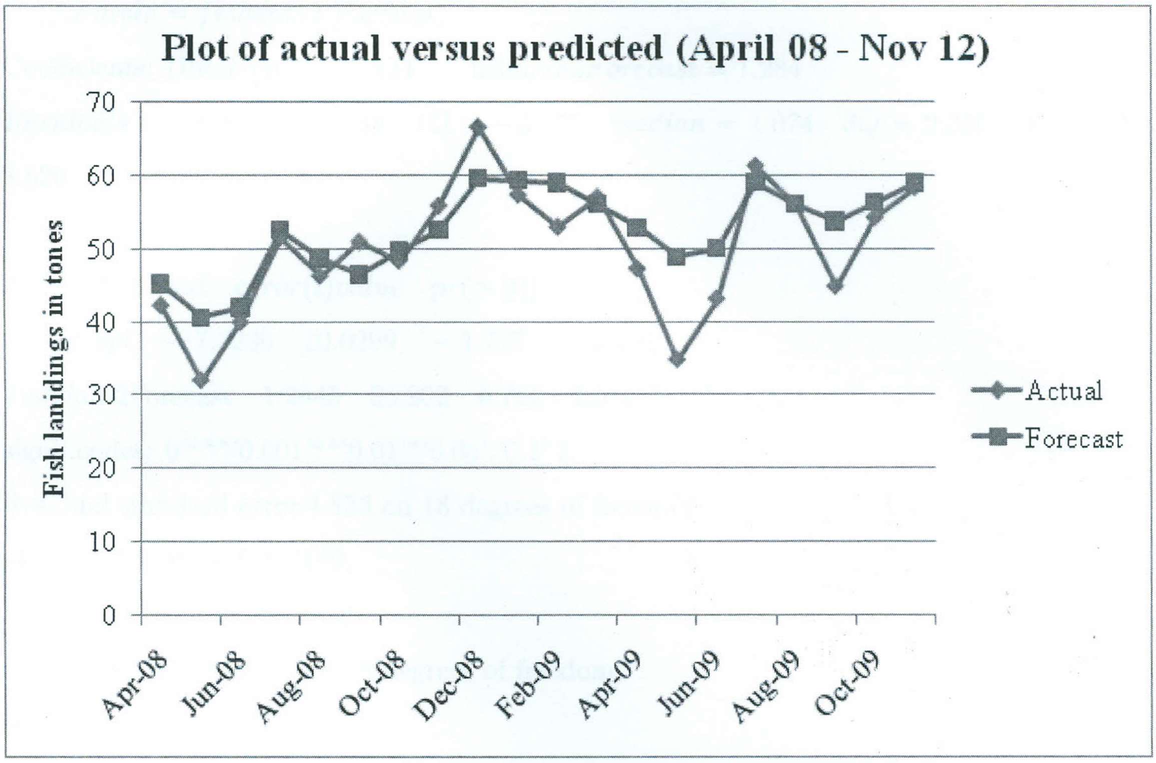


Figure 4.11: Plot of actual versus forecasted values(Apr2008-Nov2009)



The above forecasted values were superimposed on the actual values on the upper segment with a view to compare and determine the level of accuracy between the actual and predicted values obtained from the model.

4.4.2 Linear regression of the actual versus forecast

The regression analysis done using **R** gave the following results:

¿lm1-lm(fishdata2$Actualfishdata2Forecast$)

$> lm$

$Call:$

$lm(formula = fishdata2$ Forecast)

Coefficients: (Intercept)=-17.324 fishdata2$Forecast = 1.284$

$Residuals: \quad min = -10.739 \quad 1Q = -3.045 \quad median = 1.074 \quad 3Q = 2.336 \quad Max =$ 8.520

$Coefficients:$

$Estimated \quad std \quad error(t)value \quad pr(>|t|)$

$(Intercept) - 17.3236 \quad 10.0299 \quad - 1.727 \quad - 0.101$

$fishdata2$Forecast 1.2845 01.902 6.752 2.51e-06***

signif.codes: 0'***'0.001'**'0.01'*'0.05'.'0.1"1

Residual standard error:4.833 on 18 degrees of freedom

Multiple R-squared: 0.7169,

Adjusted R-squared: 0.7012

F-statistic: 45.59 on 1 and 18 degrees of freedom,

p-value: 2.507e-06.

Note: the high value of adjusted R-squared shows that the fitted values are closer to the actual ones.

Chapter 5

Research findings and Recomendations

5.1 Introduction

This chapter presents the research findings, conclusions and recommendations based on the objectives.

5.2 Conclusion

The purpose of this research was to estimate the missing values of the seasonal time series using a suitable model, we have identified the model,estimated its parameters and used it to fill the gap through forecast of the lower segment of the data. We graphed the raw data indicating the missing gaps. The autocorrelation function and the partial autocorrelation functions when plotted for the lower and upper segments indicated the SARIMA models: SARIMA$(0,1,1)(2,0,0)_{12}$ and$(0,1,2)(0,0,1)_{12}$ respectively were most suitable for the data.Forecast done from the lower segment estimated the values for

- Dec-07 54.1

- Jan-08 53.7
- Feb-08 52.8
- Mar-08 49.9

Further forecast values were obtained beyond the gap upto November-09 which were used to test the level of accuracy between the actual and predicted values.Regression analysis between the actual and the forecasted values done indicates that the predicted values are closer to the actual values signifying that the missing values estimated are better estimates.

Our research has therefore indicated that SARIMA interpolation which was developed by Box-Jenkins has provided the most suitable methods for estimating missing data.

When the missing values are quantified it is noted that about 210.4 tonnes worth of the Nile perch were not harvested due to the chaos that lasted for about four months.This had a negative impact on the economy of the region in which fishing is a major economic activity.

5.3 Suggestions

In our project we have examined time series of univariate case only. This project would have as well addressed the multivariate cases especially harvesting of other species of fish. Our project is limited to a consecutively missing values only. We suggest that non-consecutive cases are also prone to occur and should be investigated.

Our model obtained was restricted to relatively fewer observations this could have contributed to greater variability leading to a subjective model.We recommend that larger amount of data be used to enhance accuracy in modelling time series and hence use the model to make better estimates.

The project discussed the data obtained by the fishermen and recorded by the authority but did not account for unrecorded harvests i.e the quantities that were locally consumed.We therefore recommend that better mechanisms to be examined so as to take care of this significant error of omission.

References

[1] Abraham,B,"Missing observations in Time Series."Communications in statistics Theory,*10,pg,1643-1653*,1981.

[2] Akaike H."A new look at the statistical model identification.",*IEEE Transactions on Information Theory 47,pg,716-723*1974.

[3] Ansley C.F.and R.Kohn(1985)"Estimation,filtering and smoothing in state space models with incompletely specified initial conditions."*The American statistician.13,pg,1286-1316."*1985.

[4] Booth G.W and T .I Peterson(1962)."Non-Linear estimation."*Journal,American Statistical Association 57,pg,269-306.*1962

[5] Box G.E.P and G.M.Jenkins(1976)."Time Series Analysis Forecasting and Control".*Revised edition,San Fransisco.pg,21-375.*1976

[6] Box G.E.P and G.C.Tiao(1978)."Intervention analysis with applications to economic environmental problems." *Journal,American Statistical Association.vol70,70-79.*1978

[7] Gardner,G.A.C Harvey and Phillips G.D.A(1980)."An Algorithm for Exact Maximum Likelihood Estimation of Autoregressive-Moving Average Models by means of Kalman."*Aplied statistics Journal,vol29,pg,311-322.*1980

[8] Harvey A.C and R.G.Pierce(1983)."Estimating Missing Observations in Economic Time Series with Structural and Box-Jenkins Models:"*A case study,pg 299-307.*1983

[9] Holt C.C.(1957)."Forecasting trends and Seasonals by Exponentially Weighted Moving Averages."*Carnegie Institute of Tecnology,Pittsburgh.*1957.

[10] Chatfield,C.(2003)."The Analysis of Time Series:An Introduction" *(6th ed)New York, USA.John Wiley and Sons,pg 363-377.*2003

[11] Peter J, Diggle (1990)."Time series: A biostatistical Introduction" (4^{th}*ed*) *Oxford Science Publications,pg 165-202.*1990

[12] Box,G.E.P and Pierce,D.A (1970)." Distribution of autocorrelations in auotregressive-integrated moving average time series models." *Journal, American Statistical Association vol 15,1509-1546.*1970

[13] Abraham,B. and Ledolter,J.(1983)."Statistical methods for forecasting." *Wiley,New York,pg 89-167.*1983

[14] Altham,P.M.E.(1984)." Improving the precision of estimation by fitting a model." *Journal, the Royal Statistical Society,vol B 46,pg 118-129.*1984

[15] Jones,R.H.(1980)."Maximum likelihood fitting of ARMA models to time series with missing observations." *Journal, Technometrics vol22,pg389-395.*1980

[16] Lawrence,A.J and Lewis,P.A.W.(1985)." Modelling and residual analysis of nonlinear autoregressive time series in exponential variables." *Journal of the Royal Statistical Society,volB 47,pg165-202.*1985

[17] Ratowsky,D.A.(1983)."Non-linear regression modelling: a unified practical approach." *Marcel Dekker,New York,pg 123-434.*1983

[18] Velleman,P.F.(1980)."Definition and comparison of robust nonlinear data smoothing algorithms." *Journal of the American Statistical Association vol75,pg609-635.*1980

[19] Jones R.H.(1980)."Maximum Likelihood fitting of ARIMA Models to Time Series with Missing Observations". *Journal, Technometrics,vol81,pg35-45.*1980

[20] .Robinson,P.M and Dunsmuir,W.(1981)."Estimation of the Time Series Models in the prescence of Missing Data." *A journal,the American Statistical Association,vol76,pg342-349.*1981.

[21] Robinson,P.M(1985)."Testing for Serial Correlation in Regression with Missing Observations." *Journal,the Royal Statistical Society B,vol47,pg429-437.*1985.

[22] Shumway,R.H.(1982)."An Approach to time Series Smoothing and Forecasting using the EM Algorithm."*A journal of the Time Series Analysis,vol3(4),pg253-264.*1982

[23] Shively,T.S(1992)."Testing for Autoregressive Disturbances in the Time Series Regression with Missing Observations."*A journal of econometrics,vol57(1),pg,233-255.*1992

[24] Damsleth,E.(1979)."interpolating Missing Values in a Time Series."*A journal,Scand J Statisticians,vol7,pg33-39.*1979.

[25] Lyung,G.M(1989)"A note on the Estimation of Missing Values in the Time Series."*Communications in Statistics simulation,vol18(2)pg459-465.*1989.

[26] Cryer,J.D. and Chan K.S.(2008)"Time series Analysis with aplication in R."*(2nd Ed),Springer, New York,ISBN-10:pg491.*2008.

[27] Brockwell, p. and R,Davis(2002)"Introduction to time series and forecasting,2nded" *Springer,New York.pg469-478.*2002.

[28] Burnham,K.P.,and D.R.Anderson(2002)"Model selection and multi-model inference: a practical information-theoretic approach,2nded" *Springer,New York.pg488-544.*2002.

[29] Ljung,G.,and G. Box(1978)"On a measure of lack of fit in time series models" *Biometrika,vol65pg297-303.*1978.