Analysis of Seasonal Time Series with Missing Observations of Fish Harvesting in Lake Victoria, Kenya

Otwande Andrea [1] Ojung'a Okoth Samson[2] Otulo Wandera Cyrilus[1] and Fred Onyango[2]
1 Rongo University College, Rongo, Kenya
2 School of Mathematics, Applied Statistics and Actuarial science, Maseno University, Kenya
E-mail of the corresponding author:mcojijosamsa@gmail.com

ABSTRACT

Time series is a measured observation recorded with time. 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. The most commonly used approaches to estimate missing values 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 paper 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 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 analyzed 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 the strong positive relationship indicating high level of accuracy on predictability of the model used.

Keywords: Stationary and Stochastic process, Auto-covariance and Autocorrelation Function, Partial Autocorrelation Function, correlogram, Moving Average process, Autoregressive process.

1 Introduction

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 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 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.

1.1 Basic Concepts and Notation

1.1.1 Stationary 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 ofX_{t1+h}, X_{t2+h}, .., X_{tn+h} for all ($t_i \in\Re$).Hence the mean and variance if they exist do not change over time. 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,$E(Xt) = \mu < \infty$ for all $t \in \Re$

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

Where μ is a real number and δ (h) is the autocovariance function for lag of h.The transformation involves differencing the data with the given series X_t to create the new series

$$\Delta X_t = X_t - X_{t-1} \qquad (1.2)$$

1.1.2 Autocovariance and Autocorrelation Function

Autocorrelation 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 Xt we have the: Mean

$$E\,(X_t) = \mu \qquad (1.3)$$

and variance

$$\delta\,(0) = E\,(X_t - \mu)^2 = \delta^2 \qquad (1.4)$$

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

$$\rho(h) = \mathrm{Cov}\,(Xt, Xt + h)/\sqrt{V}\,ar(Xt)Var(Xt + h) \;= \delta(h)/\delta(o) \qquad (1.5)$$

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;

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

1.1.3 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 autocorrelation is the partial regression coefficient Φ_{hh} in the r^{th} order autoregression.

$$X_{t+h} = \Phi_{h1},\, X_{t+h-1} + \Phi_{h2}X_{t+h-2} + \ldots + \Phi_{hh}X_{\,t} + e_{t+h} \qquad (1.6)$$

Where e_{t+h} is normal error term. Multiplying the equation above by $X_{\,t+h-j}$ and taking the expectation the result is;

$$\delta\,(j) = \Phi_{h1}\delta_{(j-1)} + \Phi_{h2}\delta_{(j-2)} + \ldots + \Phi_{hh}\,\delta_{\,j-h} \qquad (1.7)$$

By deviding both sides by δ (0) we get

$$\rho\,(j) = \Phi_{h1}\delta_{(j-1)} + \Phi_{h2}\rho\,(j-2) + \ldots + \Phi_{hh}\,\rho\,(j-h) \qquad (1.8)$$

where $j \geq 1$. These correlation functions can be applicable mostly in modeling 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.

1.1.4 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 $r_{\,h} = \delta\,(h) / \delta\,(0)$

Where $\delta\,(h) = \sum_{t=1}^{N-h} \frac{(Xt-\hat{X})(Xt+1-\hat{X})}{N}$
for h = 0, 1, 2......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 behavior and properties of the time series.

1.1.5 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

$$Xt = \beta_0 e_t + \beta_1 e_{t-1} + ... + \beta_q e_{t-q} \qquad (1.9)$$

Where β_0, β_1, β_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 M A (q) which means Moving Average of order *q*. Basic model for the moving average is

$$Xt = e_t + \theta\, e_{t-1} \qquad (1.10)$$

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

$$\delta\,(h) = \mathrm{Cov}\,(X_t, X_{t+h}) \qquad (1.11)$$

1.1.6 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

$$Xt = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + ... + \Phi_p X_{t-p} + e_t \qquad (1.12)$$

Where Φ_1; Φ_2; Φ_p are parameters of autoregressive and the subscripts 1, 2,...are the orders of the autoregressive parameters which increase with increase in X_t.

The basic model of *AR* is $X_t = \Phi X_{t-1} + e_t$ (1.13)
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 equation 1.13 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 Φ (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 $\Phi B = 1 - \Phi B.....\Phi_p B_p$.

1.1.7 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

$$Xt = \sum_{T=1}^{P} \Phi t X_{t-1} + e_t - \sum_{T=1}^{P} \theta_j e_{t-j} \qquad (1.14)$$

Thus

$$X_t = \Phi_1 + \Phi_2 X_{t-2} + ... + \Phi_p X_{t-p} - \theta_1 e_{t-1} + \theta_2 e_{t-2} + ...\theta_q e_{t-q} \qquad (1.15)$$

When simplified using backward shift operator $B_j X_t = X_{t-j}$ We get

$$\Phi\,(B)\, X_t = \theta\,(B)\, e_t \qquad (1.16)$$

where $\theta\,(B) = 1 - \theta_1 B....\theta qBq$ and $\Phi\,(B) = 1 - \Phi_1 B...\Phi_p B_p$ are polynomials of degree *p* and *q* in B and the process is ARMA (p, q) To attain stationary for this model the equation Φ(B) has its roots lying outside the unit circle and θ(B) = 0 must lie outside the unit circle for the process to be invertible.

1.1.8 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. Consider the Autoregressive Moving Average of order (p, q) given by:-

$$Xt = \Phi_1 X_{t-1} + ... + \Phi p X_{t-p}\, e_t + \beta_1 e_1 + ... + \beta_q e_{t-q} \qquad (1.17)$$

Suppose Xt is non-stationary but ∇X_t is stationary.
Let $\omega_t = \nabla d\, X_t$, ω_t is stationary while X_t is non-stationary.
Where $\omega_t = \Phi_1\omega_1 + ... + \Phi_p\omega_{t-p} + e_t + \beta_1 e_1 + ... + \beta_q e_{t-q}$.
We can write using backward shift operator

$$\Phi\,(B)\,\omega_t = \Theta\,(B)\, e_t \qquad (1.18)$$

This implies
Φ (B)∇d $X_t = \Theta$ (B) e_t note $\nabla = 1 - B$, $\nabla X_t = X_t - X_{t-1}$, $(1 - B) X_t = X_t - (B) X_t$.
Thus

$$\Phi B(1 - B)d X_t = \Theta (B) e_t \quad (1.19)$$

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

1.1.9 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

$$\varphi p (B)\Phi P (Bs)\nabla d\nabla D$$
$$s Xt = \theta q (B)\Theta Q(Bs) e_t \quad (1.20)$$

where the subscripts θpΦP θqΘQ 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. ∇1∇ = Xt − Xt1 = (1 − B) Xt. Thus the final differenced stationary time series is not only from simple differencing to remove the trend but also seasonal ∇s to remove seasonality.

2 Literature Review

In order to remove non-stationary sources of variation and fit stationary models Box and Jenkins in 1976 [3] recommended the extension to Autoregressive Moving Average process which deals with stationary process through differencing. Richard H. Jones in 1980 came up with a method that involved calculating exact likelihood function of stationary Autoregressive Moving Average based on Akaike's Markovian [1] 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 [5]and G.D.A Phillips[4]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 [5] 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 [2] 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[7] came up with the expression for the likelihood function of parameters in Autoregressive integrated 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 focuses on forecasting approach and the recursive linear estimators obtained when the minimum mean square error are optimal. In 1997 Albert Luceno extended Lyung's [7] 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 Chatfield6 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 deals with stationary process through differencing. Richard H. Jones in 1980 came up with a method that involved calculating exact likelihood function of stationary Autoregressive Moving Average based on Akaike’s Markovian [1] 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 [5]and G.D. A Phillips[4]came up with an al-gorithm 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 on-deterministic(stochastic). In 1984, A.C.Harvey and R.G. Pierce[5] 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 [2] 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[7] came up with the expression for the likelihood function of parameters in Autoregressive integrated 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 focuses on forecasting approach and the recursive linear estimators obtained when the minimum mean square error are optimal. In 1997 Albert Luceno extended Lyung’s [7] 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[6] 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 could be used to convert non-stationary time series to stationary model by use of seasonal autoregressive moving average.

3 Methodology

The preliminary stages of this study focused on the use of Box-Jenkins Autoregressive Integrated Moving Average model to identify the most suitable model for the data obtained, 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.

4 Model building

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

The 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, ..., n) is a graph of the sample autocorrelation coefficient τ against the corresponding lags h. Each τ is defined as:

$$\tau = \frac{g_h}{g_o} = \frac{\sum_{t=h+1}^{n}(x_t-\hat{x})(x_{t-h}-\hat{x})}{\sum_{t=t}^{n}(x_t-\hat{x})(x_t-\hat{x})} \tag{4.1}$$

The level of individual departure of τ_h is checked within the limits of ±√ 2. Also the pattern *n* 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 M A (q) is a suitable model and for AR process the autocorrelations decaying exponentially is observed by partial autocorrelation function which has a cut of for an underlying process at $\varphi_{hh} = 0 \ \forall \ h > p$ According to Chartfield [6] 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

$$\varphi_p(B)\Theta_P(B^s)w_t = \theta_q(B)\Theta(B^s)e_t \tag{4.2}$$

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 τ_h = s, 2s where s is the periods.

This identified model can then be compared with theoretical patterns of known models. In summary the data is AR (p) if its ACF will decline steadily, or follow a damped cycle and P ACF will cut suddenly after *p* lags. It is a MA (*q*) if its ACF will cut of 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/P ACF estimates from a single sample. The minimum sample size generally recommended for the SARIM A model fitting is 50 observations[6]. And as the sample size become larger ACF/P ACF estimates tend to lower variability hence better approximation of the underlying process. However when the sample size is small then the interpretation of ACF/P ACF 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 (AIC) 1 and the small sample bias corrected equivalent AICc is used. Bayesian Information Criterion (BIC) can as well be used. AIC/AICc selection of the model involves estimation by maximum likelihood methods of a set of model candidates. The model candidates will then have their AIC/AICc values determined and the model candidates with minimum AIC/AICc is then selected as the model that is closest to the statistical process generating the data. AIC is calculated as

AIC = −2*ln* (L) + 2r where *ln* (L) is the log likelihood of the model and r = p + q + P + Q + 1

AICc = 2*ln* (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 = −2*ln* (L) + r + r*ln*N 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.

4.2 Model Estimation

To estimate the model parameters for ARMA model we shall consider an AR process of order 1 given as:

$$(X_t - \mu) = \alpha\,(X_{t-1} - \mu) + \varepsilon_t \tag{4.3}$$

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 = the mean value of each X_t hence we estimate $\hat{\mu} = \hat{x}$ the sample mean of the data.

α = the first autocorrelation of (X_t). So we estimate it by $\hat{\alpha} = \tau_1$, the first sample autocorrelation coefficient.

Given μ and α we can construct residuals $\varepsilon_t = (X_t - \mu) - \alpha\,(X_{t-1} - \mu)$; t = 2, 3, ..., n and n ε^2 the estimate σ^2 = var (ε_t) by the residual mean square

S (μ, α) + n (X n−1 t=2 t − μ) − α(Xt−1 − μ)2; to obtain least μ and α squares we differentiate S (μ; α) δS = 2 n (X (X δα t=2 t − μ) − α(Xt−1 − μ)(α − 1) = 2(α − 1)[n t=2 t − αXt−1) + (α − 1)(n − 1)μ] δS = −2 n (X δα t =2 t − μ) − α(Xt1 − μ)(Xt−1 − μ)

summary the data is AR(p) if its ACF will decline steadily, or follow a damped cycle and P ACF will cut suddenly after p lags. It is a M A(q) if its ACF will cut of suddenly after q lags and P ACF 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/P ACF estimates from a single sample. The minimum sample size generally recommended for the SARIM A model fitting is 50 observations6. And as the sample size become larger ACF/P ACF estimates tend to lower variability hence better approximation of the underlying process. However when the sample size is small then the interpretation of ACF/P ACF 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(AIC) 1 and the small sample bias corrected equivalent AICc is used. Bayesian Information Criterion (BIC)can as well be used. AIC/AICc selection of the model involves estimation by maximum likelihood methods of a set of model candidates. The model candidates will then have their AIC/AICc values determined and the model candidates with minimum AIC/AICc is then selected as the model that is closest to the statistical process generating the data. AIC is calculated as

$\mathrm{AIC} = -2\ln(L) + 2r$ where $\ln(L)$is the loglikelihood of the model and $r = p + q + P + Q + 1$

$\mathrm{AICc} = 2\ln(L) + 2r + 2r(r + 1)/(n - r - 1)$ where $n = N - D - d$ is the number of observations used to fit the model. And $\mathrm{BIC} = -2\ln(L) + r + r\ln N$ 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.

5 Discussion of Results

5.1 Data analysis on segment1

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

Figure 1: Monthly Lake Victoria Landing in tonnes 2001-2012

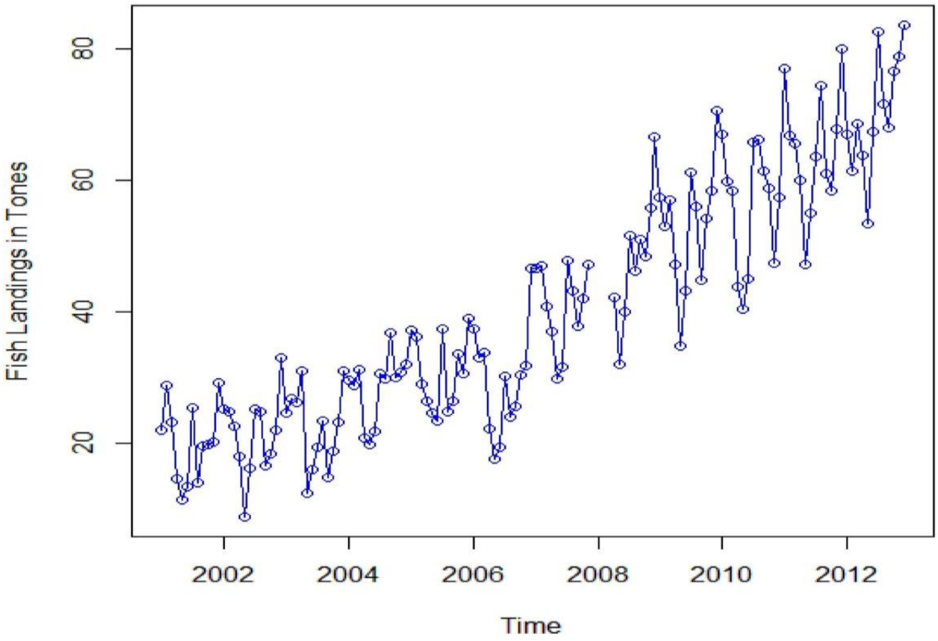
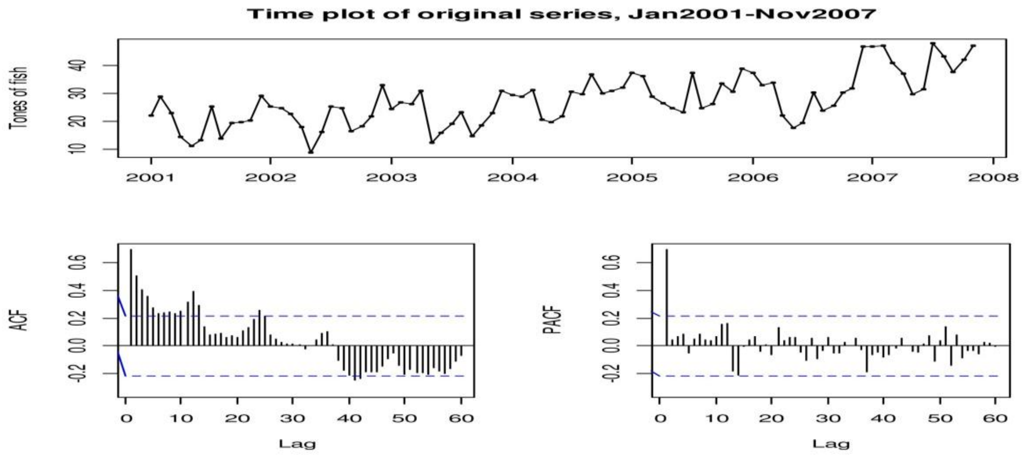


Figure 2: Time plot of the data between January 2001 and November 2007 (with ACF and PACF)

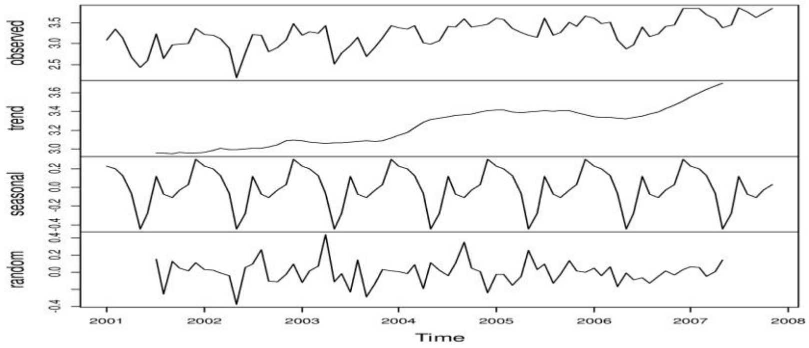


The above time plot is for the original data from January 2001 to December 2012 with entries for four months and for data to be analysed well it is advisable to divide it into segment1 (from Jan 2001 to Nov2007)and segment 2(from Apr2008 to Dec2012).

5.1.1 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 P ACF .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 stationary and should be differenced.

Figure 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.

Figure 4: Differencing the Series to achieve stationarity

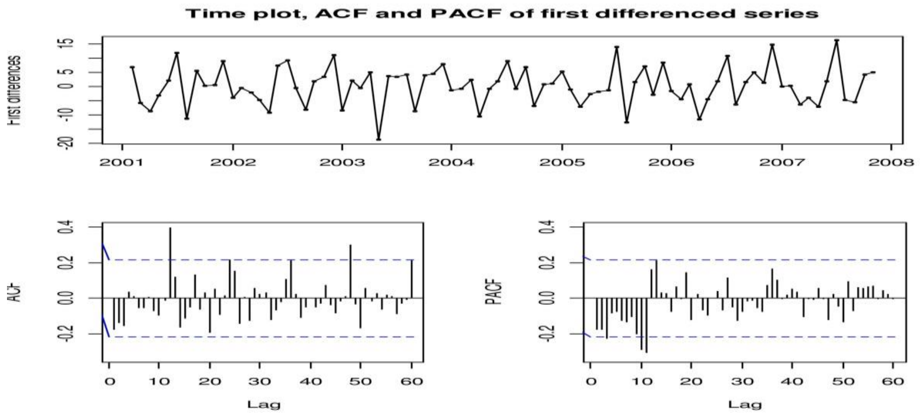
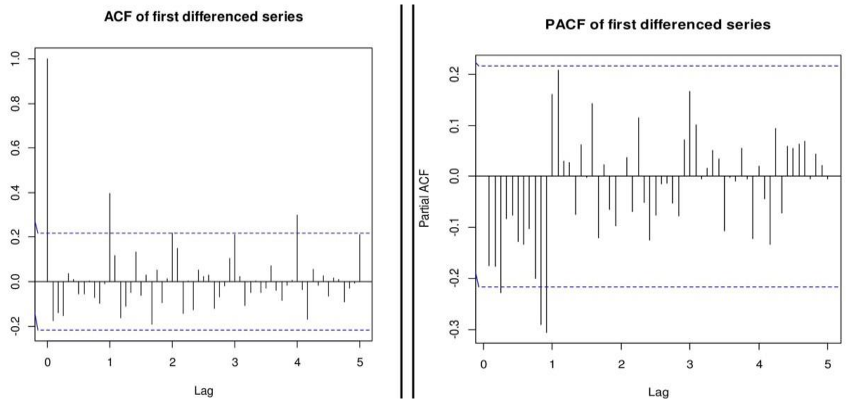


Figure 5: First Differencing the Series



5.1.2 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 ARIM A 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

5.1.3 Seasonal Differencing

A useful R function ndiffs () is used to determine whether seasonal differencing is required. Results from R: ndiffs (fishdata1ts) [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 SARIM A (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 P ACF of the differenced series, we find that either:

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

We therefore have two candidate models:

- SARIMA (0, 1, 1) ∗ (0, 0, 1)12
- SARIMA (0, 1, 1) ∗ (2, 0, 0)12

1. Model [1]: SARIMA (0, 1, 1) ∗ (0, 0, 1)12

Fishdata1ts-sarima1 Series: fishdata1ts
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 estimated as 31.87 Log likelihood= -259.81
AIC=525.63 BIC = 532.85

2. Model [2]: SARIMA (0, 1, 1) ∗ (2, 0, 0)12
fishdata1ts-sarima2 Series:fishdata1ts
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 estimated as 25.08 log likelihood= -252:58
AIC= 513.16 BIC = 522.79

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

5.1.4 Decision

Based on the results shown above, we entertain the second model; SARIM A(0, 1, 1)∗(2, 0, 0)12 since it has smaller AIC and BIC values compared to the first SARIM A(0,1,1)* (0,0,1)12.

5.2 Data analysis on Segment2

5.2.1 SARIMA Model

The suitable model identified by the automatic procedure was ARIM A(0, 1, 2) ∗ (1, 0, 0)12. The results from R are as follows: auto.arima (fishdata2ts)
Series: fishshdata2ts
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 estimated 48.65 log likelihood= -190.37 AIC= 388.74 and BIC= 397.12

5.3 Forecast

5.3.1 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 6: Forecast for ARIMA (0,1,1)(0,2,2)12

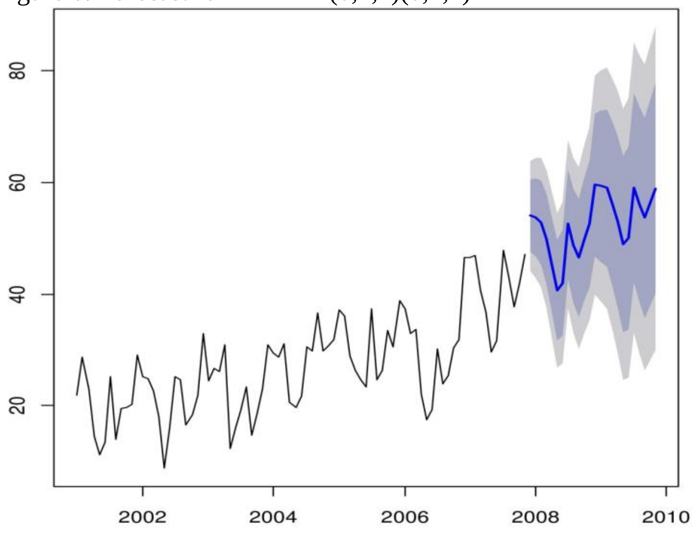
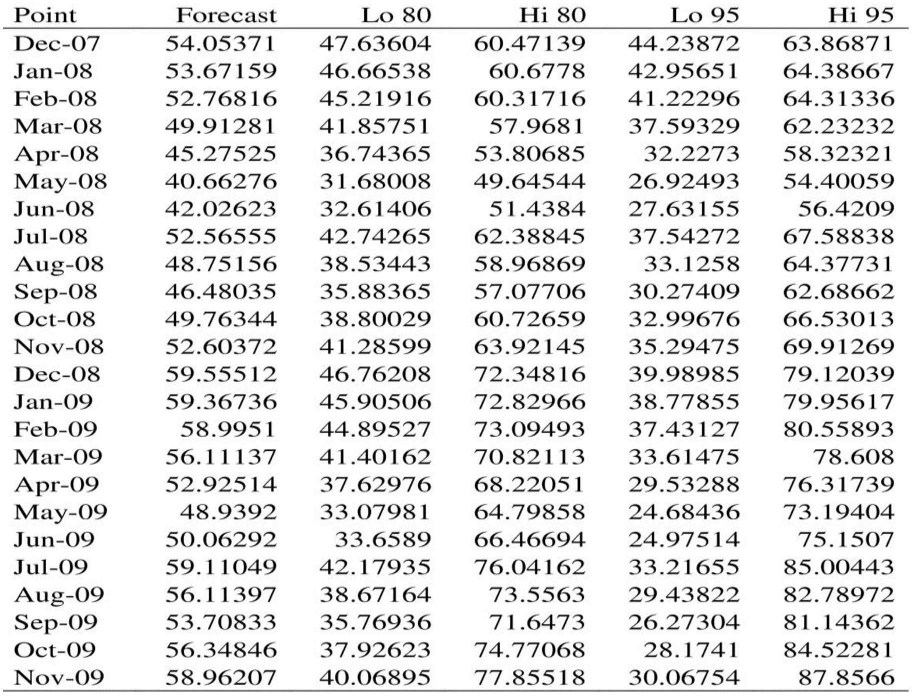


Figure 7: Forecasted Values



Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Dec-07	54.05371	47.63604	60.47139	44.23872	63.86871
Jan-08	53.67159	46.66538	60.6778	42.95651	64.38667
Feb-08	52.76816	45.21916	60.31716	41.22296	64.31336
Mar-08	49.91281	41.85751	57.9681	37.59329	62.23232
Apr-08	45.27525	36.74365	53.80685	32.2273	58.32321
May-08	40.66276	31.68008	49.64544	26.92493	54.40059
Jun-08	42.02623	32.61406	51.4384	27.63155	56.4209
Jul-08	52.56555	42.74265	62.38845	37.54272	67.58838
Aug-08	48.75156	38.53443	58.96869	33.1258	64.37731
Sep-08	46.48035	35.88365	57.07706	30.27409	62.68662
Oct-08	49.76344	38.80029	60.72659	32.99676	66.53013
Nov-08	52.60372	41.28599	63.92145	35.29475	69.91269
Dec-08	59.55512	46.76208	72.34816	39.98985	79.12039
Jan-09	59.36736	45.90506	72.82966	38.77855	79.95617
Feb-09	58.9951	44.89527	73.09493	37.43127	80.55893
Mar-09	56.11137	41.40162	70.82113	33.61475	78.608
Apr-09	52.92514	37.62976	68.22051	29.53288	76.31739
May-09	48.9392	33.07981	64.79858	24.68436	73.19404
Jun-09	50.06292	33.6589	66.46694	24.97514	75.1507
Jul-09	59.11049	42.17935	76.04162	33.21655	85.00443
Aug-09	56.11397	38.67164	73.5563	29.43822	82.78972
Sep-09	53.70833	35.76936	71.6473	26.27304	81.14362
Oct-09	56.34846	37.92623	74.77068	28.1741	84.52281
Nov-09	58.96207	40.06895	77.85518	30.06754	87.8566

Figure 8: Plot of forecasted values only

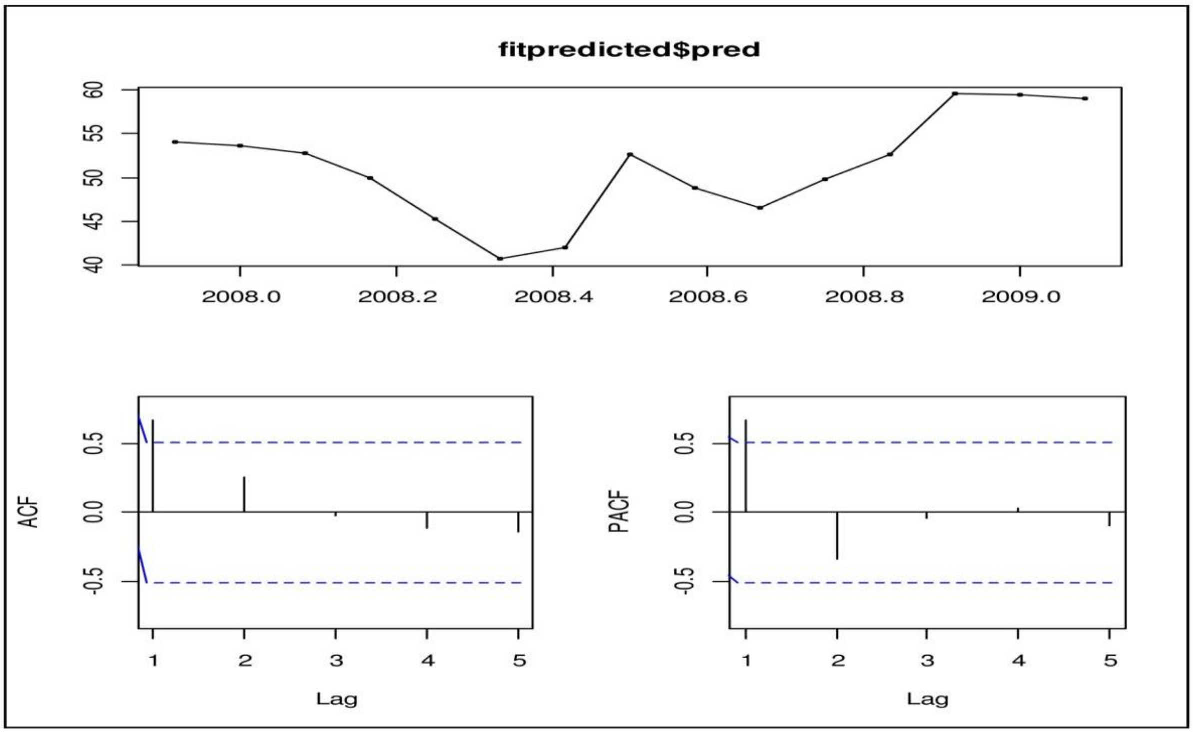
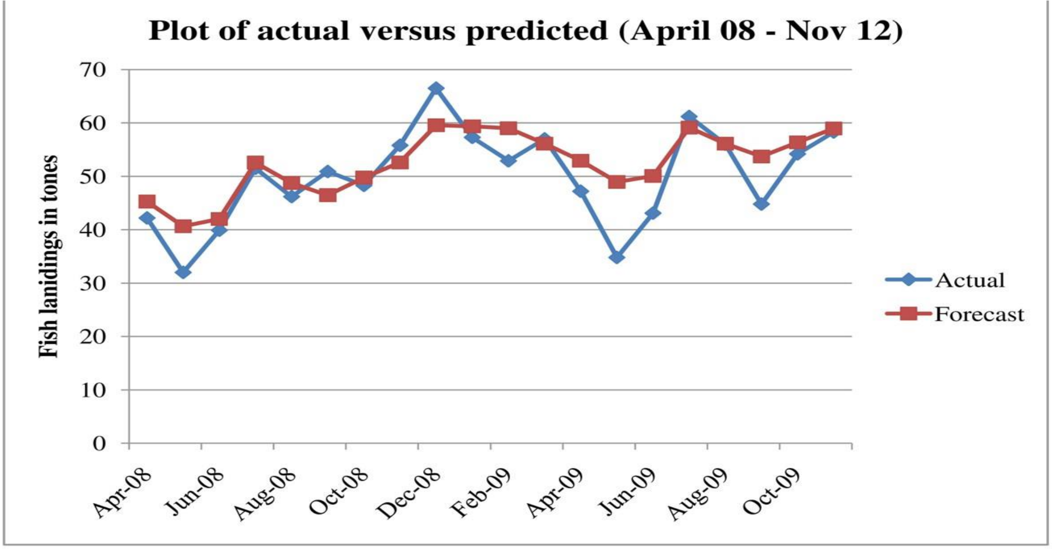


Figure 9: Plot of actual versus forecasted values for the period Apr 2008 to Nov 2009



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.

6 Conclusion and Recommendation

6.1 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-2007- 54.1
- Jan-2008 -53.7
- Feb-2008 -52.8
- Mar-2008 -49.9

Further forecast values were obtained beyond the gap upto November-2009 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.

6.2 Recommendation

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 modeling 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] Akaike H.”A new look at the statistical model identification.” IEE Transactions on Information Theory 47, pg, 716-723 1974.

[2] 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.

[3] Box G.E.P and G.M.Jenkins (1976)”Time Series Analysis Forecasting and Control”. Revised edition, an Fransisco.pg, 21-375.1976

[4] Gardner, G.A.C Harvey and Phillips G.D.A (1980).An Algorithm for Exact Maxi- mum Likelihood Estimation of Autoregressive-Moving Average Models by means of Kalman.”Applied statistics Journal, vol29, pg, 311-322.1980

[5] 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

[6] Chatfield, C. (2003).”The Analysis of Time series: An Introduction” (6th Ed) New York, USA.John Wiley and Sons, pg 363-377.2003

[7] Lyung, G.M (1989)”A note on the Estimation of Missing Values in the Time Series.” Communications in Statistics simulation, vol 1 8(2) pg459-465.1989.