Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.3, No.12, 2013 

# Imputation of incomplete non- stationary seasonal time series data

Yodah Walter.O, Kihoro, J. M, Athiany, K.H.O, Kibunja H. W

Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture & Technology,P. O. Box 62000-00200, Nairobi, Kenya

* E-mail of the corresponding author: walterkayodah@gmail.com

**Abstract**

Missing observations in time series data is a common problem that occurs due to many reasons. In order to estimate missing observation accurately, it is necessary to select an appropriate model depending on the type and nature of the data being handled so as to obtain the best possible estimates of missing observations. The objective of the study was to examine and compare the appropriateness of Box Jenkins models and direct linear regression in imputing missing observation in non stationary seasonal time series data. The study examined Box Jenkins techniques SARIMA and ARIMA models in imputing non stationary seasonal time series specifically in situations where missing observation are encountered towards the end of the series. Besides that, direct linear regression have also been proposed in imputing missing observations when seasonality has been relaxed by rearranging the time series data in periods and grouping observations which corresponds to each other from each period together and then analyze each as a single series. From the study it was observed that it is easy to impute missing observations using direct linear regression in non-stationary time series data when seasonality has been relaxed by rearranging the data in periods compared to traditional Box Jenkins models SARIMA and ARIMA models. Also direct linear regression proved, more accurate and reliable compared to Box-Jenkins techniques. So Based on the finding, the proposed direct linear regression approach can be used in imputing missing observations for non stationary series with seasonality by first rearranging the data in periods.

**KEYWORDS:** Imputation, SARIMA, ARIMA models and Direct Linear regression (L.REG).

## Introduction

Missing values in time series data is one of the problems commonly encountered. Missing values may occur due to lack of records, item non response, machine failure to record observation during experiment, lost records among others. Several techniques may be used in computing missing values. They may be simple or complex depending on the nature of time series data being handled. The most common techniques used in imputing missing values in non stationary seasonal series as suggested in literature review mainly involve the use Box-Jenkins models.

Box-Jenkins' procedures mainly entails the model identification that is selection of appropriate model, determination of appropriate values for the parameter in the model for known data patterns, model checking and lastly forecasting future values or Back-forecasting. An Autoregressive Integrated Moving Average (ARIMA) model is one of the box-Jenkins techniques that can be fitted to non stationary series as proposed by Box-Jenkins' (1976) for non stationary series which has seasonal component, then the seasonal component can be removed by seasonal differencing and the resulting model is known as seasonal ARIMA model or SARIMA model.

This paper examines the appropriateness of Box-Jenkins approaches (SARIMA and ARIMA models) in handling non-stationary seasonal time series with missing observations. Besides that, the paper also discusses direct linear regression approach in imputing missing observation when seasonality has been relaxed by rearranging the series in periods and the treating each period as a single series.

## Box-Jenkins Models

If the observed time series process is linear and non stationary process with seasonality then we will confine ourselves to the following Box-Jenkins models as discussed by Box and Jenkins (1976)

### i. Autoregressive (AR) models

A model of the form

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \cdots\cdots \phi_p y_{t-p} + e_t \qquad (1)$$

where $\phi_1, \phi_2 \cdots \phi_p$ is a set of finite weight parameters is called an autoregressive process of order $p$ that is

 

$AR(p)$. For first order and second order we have $y_t = \phi_1 y_{t-1} + e_t$ and $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + e_t$

Equation (1) can also be represented in the form

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) y_t = e_t \qquad \text{or} \qquad \phi(B) y_t = e_t \tag{2}$$

where $\phi(B) = 1 - \phi_1 B - \phi_2 B - .... - \phi_p B^p$ .From equation 2.10 we get $y_t = \frac{1}{\phi(B)} e_t = \phi^{-1}(B) e_t = \psi(B) e_t$ which implies that $\phi^{-1}(B) = \psi(B)$.

**ii. Moving Average (MA) models**

This is the second type of Box-Jenkins' models. The general equation for MA process is given by

$$y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \tag{3}$$

where $\theta_1, \theta_2 \cdots\cdots \theta_q$ are the parameters of order $q$ and $e_t, e_{t-1} \cdots\cdots, e_{t-q}$ are error terms.

If we define moving operator of order $q$ by

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

then the equation (3) above can be represented as

$$y_t = \theta(B) e_t. \tag{4}$$

**iii. Autoregressive Moving Average Models (ARMA models)**

This is a combination both AR and MA of order $p$ and $q$. The general equation for ARMA process is given by the

$$y_t = e_t + (\phi_1 y_{t-1} + \cdots + \phi_{t-p} y_p) - (\theta_1 y_{t-1} + \cdots + \phi_q e_{t-q}). \tag{5}$$

This can be simplified by backward shift operator B to obtain

$$\phi(B) y_t = \theta(B) e_t \tag{6}$$

such that $B^j y_{t-j} = y_{t-j}$, $\phi(B) = 1 - \phi_1 B - \cdots - \phi_P B^P$, and $\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^P$.

**iv. Autoregressive Integrated Moving Average (ARIMA) Models**

For time series with polynomial trend of degree *d,* the trend can be eliminated by considering a process $\Delta^d Y_t$ obtained by differencing. The process $X_t = \Delta^d Y_t$ is an ARMA (p q) satisfying stationary process. The original process $(Y_t)$ is said to be autoregressive integrated moving average of order $p, d, q$ process denoted by ARIMA( $p, d, q$ ). If $(y_t)$ follows an ARIMA model then we have

$$\varphi(B) = \phi_p(B)(1 - B)^d y_t = \theta_0 + \theta_q(B) e_t \tag{7}$$

where $\phi_p(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$ is an AR operator, $\theta_q(B) = 1 - \theta_1 B - \theta_2 B - \ldots - \theta_q B^q$ is an MA operator, $\phi_p$ and $\theta_q$ are polynomials of order $p$ and $q$ respectively with all roots of polynomial equations outside the unit circle. Considering the general ARIMA model outline in (7) above, it can be expressed in three explicit forms as described by Box and Jenkins (1976) as follows:

**(a) Difference equation form of the model**

Suppose we have $y_1, y_2, \ldots y_t$ as realization of time series data, where $y_t$ the current is value and $y_1, y_2, \ldots y_{t-1}$

 

are the previous values, then we can express the current value $y_t$ of the process in terms of previous values $y_1, y_2, \ldots y_{t-1}$ and previous values of random shock $e's$ .If

$\varphi(B) = \phi(B)(1-B)^d = 1 - \varphi_1 B - \varphi_2 B^2 - \ldots - \varphi_{p+d} B^{p+d}$ then the general model (5), with $\theta_0 = 0$ , can be expressed as

$$y_t = \varphi_1 y_t - 1 + \ldots + \varphi_{p+d} y_{t-p-d} - \theta_1 e_{t-1} - \ldots - \theta_q e_{t-q} + e_t. \quad (8)$$

**(b) General expression for the $\psi$ weight**

Consider $y_t = \psi(B) e_t$ If we perform operation on both sides with the generalized AR operator $\varphi(B)$ we obtain $\varphi(B) y_t = \varphi(B)\psi(B) e_t$. However, since $\varphi(B) y_t = \theta(B) a_t$ it follows that

$$\varphi(B)\psi(B) = \theta(B). \quad (9)$$

Therefore, the $\psi$ weights may be obtained by equating coefficients of $B$ in the expansion

$$(1 - \varphi_1 B - \ldots - \varphi_{p+d} B^{p+d})(1 + \psi_1 B + \psi_2 B^2 + \ldots) = (1 - \theta_1 B - \ldots - \theta_q B^q) \quad (10)$$

so the $\psi_j$ weights of the ARIMA process can be obtained recursively through the equations

$$\psi_j = \varphi_1 \psi_{j-1} + \varphi_2 \psi_{j-2} + \ldots + \varphi_{p+d} \psi_{j-p-d} - \theta_j \quad j > 0$$

with $\psi_0 = 1$, $\psi_j = 0$ for $j < 0$, and $\theta_j = 0$ for $j > q$ .

**(c) General Expression for the $\pi$ weights**

By first considering the model in terms previous $y's$ and current shocks $e's$, the model $y_t = \psi(B) e_t$ may also be written in the inverted form as

$$\psi^{-1}(B) y_t = e_t \quad \text{or} \quad \pi(B) y_t = (1 - \sum_{j=1}^{\infty} \pi_j B^j) y_t = e_t. \quad (11)$$

$$\text{so} \quad y_t = \pi_1 y_{t-1} + \pi_2 y_{t-2} + \ldots + a_t$$

and the $\pi(B)$ must converge on or within the unit circle since (11) is invertible. For the general ARIMA model, we can obtain $\pi$ weights by substituting (11) in

$$\varphi(B) y_t = \theta(B) e_t$$

so as to obtain

$$\varphi(B) y_t = \theta(B)\pi(B) y_t.$$

Equating coefficients of $B$ in

$$\varphi(B) = \theta(B)\pi(B). \quad (12)$$

we can get the $\pi$ weights that is

$$(1 - \varphi_1 B - \ldots - \varphi_{p+d} B^{p+d}) = (1 - \theta_1 B - \ldots \theta_q B^q) \times (1 - \pi_1 B - \pi_2 B^2 - \ldots). \quad (13)$$

Thus, the $\pi_j$ weights of the ARIMA process can be determined recursively through

$$\pi_j = \theta_1 \pi_{j-1} + \theta_2 \pi_{j-2} + \ldots + \theta_q \pi_{j-q} + \varphi_j \quad j > 0$$

with $\pi_0 = -1$, $\pi_j = 0$ for $j < 0$ and $\varphi_j = 0$ for $j > p + d$ .

**5. Seasonal Autoregressive Integrated Moving Average (SARIMA) Models**

For time series that contain seasonal periodic component which repeats itself after every s observations where (s = 12 for monthly observations). Box-Jenkins (1976) have generalize ARIMA model to deal with seasonality and define a general multiplicative Seasonal ARIMA model (SARIMA) as

 

$$\phi_p(B)\Phi_p(B^S)w_t = \theta_q(B)\Theta_Q(B^S)e_t \quad (14)$$

where $w_t$ is the differenced stationary series, $\phi_p, \Phi_P, \theta_q$, and $\Theta_Q$ are polynomials of order $p, P, q, Q$ respectively, $e_t$ is the random process and

$$w_t = \nabla^d \nabla^D y_t . \quad (15)$$

For $P = 1$ the term $\Phi_p(B^S) = 1 = cons\tan t \times B^s$ m which imply that $w_t$ depends on $w_{t-s}$ since $B^s w_t = w_{t-s}$ where $w_t$ are formed from the original series $y_t$ by simple differencing and also by seasonal differencing $\nabla_s$ to remove seasonality for instance if $d = D = 1$ and $s = 12$, then

$$w_t = \nabla\nabla_{12} y_t = \nabla_{12} y_t - \nabla_{12} y_{t-1}$$
$$= (y_t - y_{t-12}) - (y_{t-1} - y_{t-13})$$

So equation (12) is said to be SARIMA model where $d$ and $D$ need not to exceed one.

## 1. Methodology

By first identifying the order, parameter estimation and model checking as described by Box-Jenkins (1976) we preceded to forecast using SARIMA model as follows:

### *1. Forecasting using SARIMA model*

Consider the time series function $Y_t$ where $t = 1,2.....n$ from with four components: trend, cyclic, seasonal and random respectively with realization $y_1, y_2, ... y_n, y_{n+1}, y_{t+2}, ... y_{n+m}$. Suppose we have the missing observations ($y_{n+1}, ... y_{n+m}$) within the data set, we fit SARIMA model to the observed values ($y_1, y_2 ..... y_n$) by first Identifying the order, Estimation of Parameters, model checking, and lastly Forecasting or Back-Forecasting based on the three approaches for forecasting as describe by Box and Jenkins (1976) as follows:

### a. Difference equation form

Assuming we have an SARIMA (1, 0, 0) $(0, 1, 1)_s$ model and (s=12) then from equation (2.7) we have

$$(1 - \alpha B)W_t = (1 + \theta B^{12})e_t$$

where $W_t = \nabla_{12} y_t$ then

$$y_t = y_{t-12} + \phi(y_{t-1} - y_{t-12-1}) + e_t + \theta e_{t-12}$$

then we find $\hat{y}(n,1) = y_{n-11} + \phi(y_n - y_{n-12}) + \theta e_{n-11}$ and $\hat{y}(n,2) = x_{n-10} + \phi[\hat{y}(n,1) - y_{n-11})] + \theta e_{n-10}$ Chatfield (2003). Forecast for future values will be calculated recursively in the same way.

**Example**

Considering U.S Birth data which has SARIMA (0, 1, 1) × $(0, 1, 1)_s$ model, one-step ahead can be obtained as follows: From the model we have

$$(1 - \theta B)(1 - \theta B^{12})e_t = (1 - \Theta B^{12} - \theta B + \theta\Theta B^{13})e_t$$

$$y_t = y_{t-1} + y_{t-12} - y_{t-13} + e_t - \theta_{t-1} - \Theta e_{t-12} + \theta\Theta e_{t-13}$$

where $\theta = -0.62$ and $\Theta = -0.801$ are parameters estimates

$$y_t = y_{t-1} + y_{t-12} - y_{t-13} + e_t + 0.62e_{t-1} + 0.801e_{t-12} + 0.49e_{t-13}$$

Taking *t* as the origin, one step ahead forecast can be obtained as:-

$$y_{t+1} = y_t + y_{t-11} - y_{t-12} + e_{t+1} + 0.62e_t + 0.801e_{t-11} + 0.49e_{t-12}.$$

### b. Using $\psi$ weights

The weights $\psi$ in the equation $Y_t = \psi(B)e_t$ is calculated and then used in computing forecast errors. Since $y_{n+k} = e_{n+k} + \psi e_{n+k} + .....$ it is clear that $\hat{y}(n,k) = \sum_{j=0}^{\infty} \psi_{k+j} e_{n-j}$ .The $k-$steps ahead forecast error is given by

 

$e_{n+k}+\psi e_{n+k}+.....+\psi_{k-1}e_{n+1}$ and the variance $k-$steps a head error is $(1+\psi^2+\psi^2)\sigma_e{}^2$.

**c. Using $\pi$ weights**

In this case the weights $\pi$ as defined in the equation $\pi(B)Y_t = e_t$ and since

$$Y_{n+k} = \pi Y_{n+k-1} + .... + \pi_k Y_n + .... + e_{n+k}$$

and $$\hat{y}(n,k) = \pi_1 \hat{y}(n,k-1) + \pi_2 \hat{y}(n,k-2) + ....\pi_k Y_n + \pi_{k+1} y_{n-1} + ....$$

The forecast can be computed recursively replacing future values with predicted values.

Using the above procedure, we adopted one step ahead forecast in order to estimate each missing values at a time and this implies that if we have ***k*** missing values then we will also have ***k*** re-estimation of parameters. The SARIMA forecasting Steps are as follows:

i. Forecast the first missing value $y_{m+1}$ using the non missing observations before $y_{m+1}$.

ii. Forecast the second missing observation $y_{m+2}$ using the non missing observations before $y_{m+2}$ with the already forecasted $y_{m+1}$ value included as non missing observation.

iii. Forecast the third missing observation $y_{m+3}$ using the non missing observations before $y_{m+3}$ with the already forecasted $y_{m+1}, y_{m+2}$ value included as non missing observation.

iv. The same procedure is repeated for the remaining missing observations (values) $y_{m+4}, y_{m+5}, ... y_{m+n}$.

v. Suppose we have very few observations before the first missing observation, we back-forecast following steps (i) to (iv) above.

**2. Rearranging the series**

Before imputing missing values using Box- Jenkins ARIMA model and direct linear regression, we first rearrange the original time series data in periods $P_n$ mainly to eliminate the seasonal component. The length of the period may be obtained by first plotting the series to examine cyclic pattern of the series. We then approximate the time gap between two successive troughs or crest of the cyclic component which we later use as the length of the period to be used in rearrangement of the series. Generally rearrangement of the series may be done as follows: Consider time series data with ***N*** observations, we may have $P = (P_1, P_2, P_3, .., P_n)$ periods where

$$\begin{aligned} P_1 &= (y_1, y_2, y_3, y_4 ..., y_{m-1}, y_m) \\ P_2 &= (y_{m+1}, y_{m+2}, ..., y_{2m-1}, y_{2m}) \\ P_3 &= (y_{2m+1}, y_{2m+2}, ..., y_{3m-1}, y_{3m}) \\ &\vdots \\ P_n &= (y_{\{(n-1)\times m\}+1}, y_{\{(n-1)\times m\}+2}, ..., y_{mn-1}, y_{mn}) \end{aligned}$$

in this case, $n$ represent the number of periods within the original time series while $m$ represent the number of observations in each period. The whole idea appears as shown in the array below on rearranging.

| $Y_t$ | $X_1$ | $X_2$ | ... | $X_{m-1}$ | $X_m$ |
|---|---|---|---|---|---|
| $y_1$ | $x_1$ | $x_2$ | ... | $x_{m-1}$ | $x_m$ |
| $y_2$ | $x_{m+1}$ | $x_{m+1}$ | ... | $x_{2m-1}$ | $x_{2m}$ |
| $y_3$ | . | | | . | . |
| $y_4$ | . | | | . | . |
| $y_5$ | . | | | . | . |
| $y_6$ | $x_{\{(n-2)m\}+1}$ | $x_{\{(n-2)m\}+2}$ | ... | $x_{\{(n-2)m\}-1}$ | $x_{(n-2)m}$ |
| $y_7$ | $x_{\{(n-1)m\}+1}$ | $x_{\{(n-1)m\}+2}$ | ... | $x_{nm-1}$ | $x_{nm}$ |
| $y_8$ | | | | | |
| $y_9$ | | | | | |
| . | | | | | |
| . | | | | | |

 

.

$y_{N-1}$

$y_N$

From the array of the rearranged series above $Y_t$ represent the original time series data while $X_1, X_2, ..., X_{m-1}$ and $X_m$ are formed by rearranging the data by seasons. Since we are interested in imputing missing values in time series data with both seasonal and non seasonal component, this arrangement removes the short circles of the data and assumes that the model of the non-seasonal component of the data remains unchanged even if the seasonal component has been removed by rearranging the data that is, suppose the original series has the model SARIMA(p, d, q) (P,D,Q) then after rearranging, the seasonal part disappear while the non seasonal part remains unchanged. So the model reduces to ARIMA (p, d, q) for the series $X_1, X_2, ..., X_{m-1}$ and $X_m$ .This assumption is only applicable if the formed re-arranged series by seasons are too short for the normal fitting of ARIMA model procedures. On the other hand if the data formed from the rearrangement of the original series is long enough, then we fit ARIMA model to each series $X_1, X_2, ..., X_{m-1}$ and $X_m$ as described by Box-Jenkins procedures which involve: model identification, parameter estimation and diagnostic checking as already been illustrated in the previous sections.

## 3 Forecasting using ARIMA model

Starting with $X_1$ series, we proceed to estimate missing values $y_{m+1}, y_{m+2}, y_{m+3}, ...$ using Box-Jenkins approaches as follows: Consider Minimum mean Square error forecast for $\hat{y}_t(k)$ of $y_{t+k}$ which is given by conditional expectation $\hat{y}_t(k) = E(y_{t+k} \mid y_t, y_{t-1}, ...)$, the actual forecast can be calculated directly in either three different ways as follows:

**a. Forecast from Difference equation**

From difference equation form as already been discussed in the previous sections, Suppose $\varphi(B) = \phi(B)(1-B)^d = (1 - \varphi_1 B - \varphi_2 B^2 - ... - \varphi_{p+d} B^{p+d})$, then the general ARIMA model in (7) can be written using difference equation form as follows

$$(1 - \varphi_1 B - \varphi_2 B^2 - ... - \varphi_{p+d} B^{p+d}) y_t = (1 - \theta_1 B - .... - \theta_q B^q) e_t \quad (16)$$

Which if we take the conditional expectation at time $t$, we obtain

$$\hat{y}_t(k) = \varphi_1[y_{t+k-1}] + ... + \varphi_{p+d}[y_{t+k-p-d}] - \theta_1[e_{t+k-1}] - ... - \theta_q[e_{t+k-q}] + [e_{t+k}] \quad (17)$$

**b. Forecast in terms of $\psi$ weights**

Using the conditional expectation in equation we obtain

$$\hat{y}_t(k) = [e_{t+k}] + \psi_1[e_{t+l-1}] + ... + \psi_{k-1}[e_{t+1}] + \psi_1[e_t] + \psi_{k+1}[e_{t-1}] + ... \quad (18)$$

Alternatively,

$$\begin{aligned} \hat{y}_t(l) &= [e_{t+h}] + \psi_1[e_{t+h-1}] + ... + \psi_{t+k-h-1}[e_{k+1}] + c_h(t+l-h) \\ &= [e_{t+k}] + \psi_1[e_{t+k-1}] + ... + \psi_{l-1}[e_{t+1}] + c_t(k). \end{aligned} \quad (19)$$

**c. Forecast in terms of $\pi$ weights**

Finally taking the conditional expectation in equation of ARIMAmodel we get

$$\hat{y}_{t(}(k) = \sum_{j=1}^{\infty} \pi_j[y_{t+k-j}] + [e_{t+k}]. \quad (20)$$

From the above forecast procedures we can obtain the missing values as illustrated in the following steps:

 

i. Forecast the first missing value $x_{m+1}$ using the non missing observations before $x_{m+1}$.

ii. Forecast the second missing observation $x_{m+2}$ using the non missing observations before $x_{m+2}$ with the already forecasted $x_{m+1}$ value included as non missing observation.

iii. Forecast the third missing observation $x_{m+3}$ using the non missing observations before $x_{m+3}$ with the already forecasted $x_{m+1}, x_{m+2}$ value included as non missing observation.

iv. Repeat the same procedure for the remaining missing observations $x_{m+4}, x_{m+5}, \ldots x_{m+n}$

v. Suppose we have very few observations before the first missing observation, we back-forecast following steps (i) to (iv) above.

Again the same steps (i) up to (iv) above will be repeated for the remaining series $X_2, X_3, \ldots, X_{m-1}, X_m$.

## 4. Imputing missing observation using direct linear regression

Considering the newly formed series $X_1, X_2, X_3, \ldots, X_{m-1}, X_m$ resulting from the rearranged original data, we first check whether there exists autocorrelation between $X_1, X_2, \ldots, X_{m-1}, X_m$ and proceed by Regressing $X_1$ on $X_2$, $X_2$ on $X_3$, $X_3$ on $X_4$ $X_4$ on $X_5, \ldots, X_{m-1}$ on $X_m$ which will yield to the following equations

$$X_1 = a_1 X_2 + b_1 + e_1,$$
$$X_2 = a_2 X_3 + b_2 + e_2$$
$$\vdots$$
$$X_{m-2} = a_{m-2} X_{m-1} + b_{m-2} + e_{m-2}$$
$$X_{m-1} = a_{i-1} X_m + b_{m-1} + e_{m-1}.$$

where $a, b, e$ are constants. Finally we can use the above regression equations to impute missing observations. For instance suppose we have time series data $(Y_t)$ with $N$=28 observations, period ( $s = 4$ ) and $y_8^m, y_{10}^m, y_{15}^m, y_{17}^m, y_{18}^m, y_{19}^m, y_{20}^m$ as missing observations at random, then on rearranging $Y_t$ we have the array as shown in table below.

$$Y_t = y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8^m y_9, y_{10}^m, y_{11}, y_{12}, y_{13}, y_{14}, y_{15}^m, y_{16}, y_{17}^m, y_{18}^m, y_{19}, y_{20}^m, y_{21}, y_{22}, y_{23}, y_{24}, y_{25}, y_{26}, y_{27}, y_{28}.$$

**Table 1 Array of rearranged series from original series $Y_t$**

| $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|
| $y_1$ | $y_2$ | $y_3$ | $y_4$ |
| $y_5$ | $y_6$ | $y_7$ | $y_8^m$ |
| $y_9$ | $y_{10}^m$ | $y_{11}$ | $y_{12}$ |
| $y_{13}$ | $y_{14}$ | $y_{15}^m$ | $y_{16}$ |
| $y_{11}^m$ | $y_{18}^m$ | $y_{19}$ | $y_{20}^m$ |
| $y_{21}$ | $y_{22}$ | $y_{23}$ | $y_{24}$ |
| $y_{25}$ | $y_{26}$ | $y_{27}$ | $y_{28}$ |

where $X_1, X_2, X_3, X_4$ are new series formed from the original series $Y_t$ now if we regress $X_1$ on $X_2$, $X_2$ on $X_3$, $X_3$ on $X_4$ and we obtain the following equations

$$X_1 = a_1 X_2 + b_1 + e_1, \quad \text{(i)}$$
$$X_2 = a_2 X_3 + b_2 + e_2 \quad \text{(ii)}$$

 

$$X_3 = a_3 X_4 + b_3 + e_3, \qquad \text{(iii)}$$

Using equation (i) and (iii) we have $y_{10}^m = \frac{y_9 - b_1 - e_1}{a_1}$ and $y_8^m = \frac{y_7 - b_3 - e_3}{a_2}$. Again using equation (ii) we have $y_{18}^m = a_2 y_{19} + b_2 + e_2$. Similarly the rest of the missing observation will be imputed using the same logic while applying the regression equations above. Note this imputation procedure only works if not all the observations in the entire row of $X_1, X_2, .., X_{m-1}, X_m$ are missing. Finally the imputed missing values in $X_1, X_2, ..., X_{m-1}$ and $X_m$ are then replaced back in the original time series data $Y_t$

## 2. Data Analysis and Discussions

We begin by setting missing values at various positions at random within each data set with missing percentages say 5%, 7%, 10%, 12% and 15% and treat each of the percentage of missing values as a sample size . Since four non stationary data sets were used in the analysis, the total number of samples were twenty. The rationale was mainly to find out which method of imputation performs better as the percentage number of missing values keeps on increasing within the data set. The data set that were used in the analysis were as follows:

- Airline data (N=144): international airline passengers: monthly totals in thousands. Jan49-Dec60. Source: time series data Library or Box-Jenkins (1976).
- The U.S Births data (N=157): Monthly U.S Births in thousands Jan 1960-Feb 1973 Source time series Library.
- Tourist data (N=228): monthly totals in thousands of world tourist visiting Kenya. Source: Kihoro (2006)
- U.K Coal consumption (N=108): quarterly totals in millions 1960-1992. Source Harvey, A. C. (2001).

Considering the following desirable properties of imputation method as suggested by Kihoro (2006)

i. Predictive accuracy: The imputed values should be very close to the actual values in order to minimize biasness.
ii. Ranking accuracy: The ordering relationship between imputed values should be the same as those of true values
iii. Distribution accuracy: This implies that, the marginal and higher order distributions of the imputed values should be essentially the same as the corresponding distributions of the true values.
iv. Estimation Accuracy: this imply that the imputation method chosen should lead to unbiased and efficient inferences for parameters of the distribution of the true values
v. Imputation plausibility: the findings of the imputation method should be values which are plausible.

the following the statistical proximity measures were used determine the similarities and dissimilarities between the imputed values and the original values for each sample size.

**a.** Product Moment Correlation Coefficient (PMCC) commonly used as a measure of similarity pattern expressed as

$$\hat{\rho}_{y\hat{y}} = \frac{\text{cov}(y, \hat{y})}{\sqrt{\text{var}(y)\,\text{var}(\hat{y})}} \qquad (21)$$

**b.** Mean Relative Euclidean Distance (MRED) which is a distance measure commonly used to measure dissimilarity (wei,(1989) and is given by

$$\hat{d}_{y\hat{y}} = \sqrt{\frac{1}{n}\sum_{n=1}^{n}\left(1 - \frac{\hat{y}_i}{y_i}\right)^2} \qquad (22)$$

**c.** Root Mean Square Error (RMSE) is also another distance measure commonly used. In this case If $\bar{d}_{y\hat{y}} = 0$ then the imputed values are very accurate. It is expressed as

$$\bar{d}_{y\hat{y}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (23)$$

**d.** Mean scaled Euclidean Distance (MSED) given as

 

$$\bar{d}_{y\hat{y}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{y_i - \hat{y}_i}{\hat{y}_i^2 + y_i^2}} = \sqrt{\frac{1}{2n}\sum_{i=1}^{n}\left(1 - \frac{2\hat{y}_i y_i}{\hat{y}_i^2 + y_i^2}\right)} \quad (24)$$

Kihoro (2006). For perfect match we expect MSED values to be zero.

**e.** Proximity measure (PROX): This is a combination of MSED and PMCC denoted by $p_{y\hat{y}}$ and takes the form

$$p_{yy} = \frac{\hat{\rho}_{y\hat{y}} - \bar{d}_{y\hat{y}} + 2}{3} \quad (25)$$

where $\bar{d}_{y\hat{y}}$ and $\hat{\rho}_{y\hat{y}}$ are MSED and PMCC values obtained from equation 4.1 and 4.4 respectively.

The values $p_{yy} = 1$ implies a perfect match of imputed values and the true values while $p_{yy} = 0$ implies that the imputed values and true values do not match at all Kihoro (2006).

## 1. SARIMA forecasting results

For each of the three data set mentioned above, seasonal ARIMA model was fitted after following box-Jenkins four steps in modeling time series and the appropriate model was obtained by choosing the model which yielded minimum AIC, and BIC. Using Box-Ljung test statistic and all the models passed the residual normality test and the finding are summarized in Table below:

**Table 2. SARIMA models used in forecasting**

| Data | Number of observation | Transformation | Model |
|---|---|---|---|
| U.S Births | 157 | Logarithm | ARIMA(0,1,1)(0,1,1) |
| Tourist series | 228 | Logarithm | ARIMA(0,1,1)(0,1,1) |
| Airline series | 144 | Logarithm | ARIMA(0,1,1)(0,1,1) |
| U.K Coal consumption series | 108 | Logarithm | ARIMA(1,1,1)(0,1,1) |

In our case, if there are $K$ - missing observations and no two or more consecutive missing observation, then we performed $K$ -re-estimation of model parameters. For instances where we have two or more consecutive missing observation, we estimate the model parameter only ones before we forecast or back-forecast the missing observations. The sample forecast for Airline data with 5% missing observation are displayed in Table 3.

## 2. ARIMA forecasting Results

For the case of ARIMA models the forecasting steps were the same as those of SARIMA model only that ARIMA models were fitted to each of the newly formed series after rearranging the data as described earlier. The newly formed series and the generated results from ARIMA forecast for sample equivalent to 5% missing observations are displayed in Table 4

## 3. Direct linear Regression Results (L.REG)

For the newly formed series which resulted from each of the data sets after rearranging, correlation between the newly formed series was examined before regressing the newly formed variables on each other after rearranging the series. Missing observation were then imputed as already been illustrated previously and the findings are also given in Table 3

 

**Table 3. Sample 1 with 5% missing observations (Airline data)**

| Position | airline | ARIMA | SARIMA | L.REG |
|---|---|---|---|---|
| 17 | 125 | 118.8800 | 132.6900 | 134.2100 |
| 75 | 267 | 269.1400 | 277.9700 | 264.0800 |
| 76 | 269 | 260.9400 | 270.0100 | 246.0000 |
| 95 | 271 | 251.3600 | 275.1500 | 270.7339 |
| 98 | 301 | 276.1500 | 308.1400 | 299.0350 |
| 102 | 422 | 415.2100 | 408.4600 | 407.1300 |
| 122 | 342 | 326.4000 | 335.3100 | 339.0400 |
| 129 | 463 | 437.4400 | 460.9500 | 468.0050 |
| 140 | 606 | 605.4300 | 628.0200 | 625.5500 |

**Empirical comparisons SARIMA, ARIMA and Linear Regression (L.REG)**

Comparing the performance of the three methods of imputing missing values using statistical measures already discussed before, Table 4 shows that, for sample size equivalent to 5% in missing observation airline passenger data, the use of MRED and RMSE indicates that (L.REG) performed better in terms of distance followed by SARIMA and ARIMA was the poorest among the three methods. In terms of preservation of data pattern (L.REG) was Again the best while ARIMA was the worst.

**Table 4. Comparison Statistics for Airline imputed values sample size 5%**

| | ARIMA | SARIMA | L.REG |
|---|---|---|---|
| **PMCC** | 0.9995 | 0.9998 | 0.9997 |
| **RMSE** | 7.3723 | 2.7893 | 1.0180 |
| **MRED** | 0.0175 | 0.0123 | 0.0025 |
| **MSED** | 0.0070 | 0.0045 | 0.0057 |
| **PROX** | 0.9975 | 0.9984 | 0.9980 |

In Table 5, the results shows that when the sample size missing observation was increased from 5% to 7% again (L.REG) was superior in terms of distance measures as can be observed by the values of RMSE, MRED and MSED.

**Table 5. Comparison Statistics for Airline imputed values sample size 7%**

| | ARIMA | SARIMA | L.REG |
|---|---|---|---|
| **PMCC** | 0.9992 | 0.9991 | 0.9998 |
| **RMSE** | 7.9958 | 12.772 | 1.8278 |
| **MRED** | 0.0330 | 0.0511 | 0.0077 |
| **MSED** | 0.0104 | 0.0111 | 0.0052 |
| **PROX** | 0.9963 | 0.9960 | 0.9982 |

Likewise the values of PMCC and PROX show that linear regression was the best in preserving data pattern. On the other hand SARIMA was the poorest in both distance as well pattern measure.

Performing similar analysis as the one indicated in table 5 and 6 for the rest of sample sizes 10%, 12%, 15% for the all data sets then by aggregating the PMCC, RMSE, MRED, MSED and PROX values, a table of ranks based on the performance of each the three method was generated and the results are indicated in table 6:

 

Now considering the overall ranking in Table 7 generated from Table 6, it is clear that in terms of preservation of both distance and pattern of the of the original data L.REG. outperformed both SARIMA

**Table 6 Ranks performance based on statistical measures for the three data sets**

| DATA | | METHOD | PMCC | Rank | RMSE | Rank | MRED | Rank | MSED | Rank | PROX | Rank |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Airline data | 1 | ARIMA | 0.9844 | 3 | 9.2912 | 3 | 0.2088 | 3 | 0.0311 | 2 | 0.9811 | 3 |
| | 2 | SARIMA | 0.9873 | 2 | 2.7113 | 2 | 0.0598 | 2 | 0.0367 | 3 | 0.9835 | 2 |
| | 3 | L-REG | 0.9942 | 1 | 1.7161 | 1 | 0.0558 | 1 | 0.0226 | 1 | 0.9905 | 1 |
| U.S Births data | 1 | ARIMA | 0.9983 | 2 | 13.938 | 3 | 0.0392 | 3 | 0.0112 | 2 | 0.9957 | 2 |
| | 2 | SARIMA | 0.9979 | 3 | 5.1030 | 2 | 0.0308 | 2 | 0.0124 | 3 | 0.9952 | 3 |
| | 3 | L-REG | 0.9989 | 1 | 4.0335 | 1 | 0.0119 | 1 | 0.0080 | 1 | 0.9970 | 1 |
| Tourist data | 1 | ARIMA | 0.9514 | 3 | 1.3997 | 3 | 0.0533 | 3 | 0.0150 | 3 | 0.9788 | 3 |
| | 2 | SARIMA | 0.9808 | 1 | 0.2220 | 2 | 0.0108 | 1 | 0.0087 | 1 | 0.9907 | 1 |
| | 3 | L-REG | 0.9515 | 2 | 0.4463 | 1 | 0.0153 | 2 | 0.0144 | 2 | 0.9790 | 2 |

and ARIMA and this suggest that Linear regression can be applied in imputing missing observation for non stationary seasonal series.

**Table 7 Overall ranks performance for the three methods**

| | PMCC | Rank | RMSE | Rank | MRED | Rank | MSED | Rank | PROX | Rank |
|---|---|---|---|---|---|---|---|---|---|---|
| **ARIMA** | 0.97803 | 3 | 8.209633 | 3 | 0.10043 | 3 | 0.01910 | 2 | 0.98520 | 3 |
| **SARIMA** | 0.98867 | 1 | 2.678767 | 2 | 0.03380 | 2 | 0.01927 | 3 | 0.98980 | 1 |
| **LREG** | 0.98153 | 2 | 2.065300 | 1 | 0.02767 | 1 | 0.01500 | 1 | 0.98883 | 2 |

## 4. Conclusion

Based on our objective we can conclude that direct linear Regression proved to be more efficient and effective if it is applied to series which has been rearranged compared to box Jenkins ARIMA and SARIMA model however it may not be applicable to all types of series thus making it inappropriate in such situations. Even though ARIMA model did not perform much compared with the other two models it can still be applied in imputing missing values where seasonality has been relaxed by rearranging data in periods. Besides that we also conclude that seasonality can also be removed by arranging the data into periods as opposed traditional method of eliminating seasonality by differencing.

## 5. Recommended areas for future research

Throughout our study, we majorly concentrated on non-stationary with seasonality; we propose the same study can be extended further for stationary series with seasonal component. We also recommend an improvement to ARIMA model specifically where newly formed series is too short for ARIMA model to be fitted instead of assuming that the model of the non-stationary part remains unchanged. Lastly we propose further study in instances where the correlation of the newly formed series is very low thus making it difficult to apply direct linear regression in imputing missing values.

## References

Abraham, B. (1981). Missing observations in time series. *Communications in statistics Theory*, 10, 1643-1653.

Akaike, H.(1973). Information theory and extension of the maximum likelihood principle, *proc.2nd international Symposium on information theory*, 267-281, Akademiai Kiado, Budapest.

 

Akaike, H.(1974). Anew look at the statistical model identification, *IEEE Transaction on Automatic control*, AC-19, 716-723.

Akaike, H.(1979). Bayesian extension of the minimum AIC procedure of Autoregressive model fitting, *Biometrika*, 66,237-242.

Akaike, H.(1979). Bayesian analysis of minimum AIC procedure, *Ann. Inst.Statist. Math*, 30 A, 9-14.

Batista, GEAPA.,and Mornad, M.C.(2003). Experimental comperition of K-nearest Neighbour and Mean or Mode Imputation Mode Imputation Methods With the internal Strategies used by C4.5 and CN2 to Treat Missing Data. Tech. Rep. 186, ICMC-USP.

Beveridge, S. (1992). Least square estimation of missing values in time series. *Communication - in statistics Theory*, 21, 3479-3496.

Brockwell, P.J. and Davis, R.A. (2002), *Introduction to time series and forecasting,* springer, New York.

Brockwell, P.J. Davis, R.A. (2006). *Time series: theory and methods.*springer, New York

Box, G.E.P, Jenkins, G.M. (1976).*Time series analysis forecasting and control*,2nd ed., San Fransico, Holden-Day.

Box, G.E.P., and Pierce, D.A (1970). Distribution of residual autocorrelation in autoregressive integrated moving average time series models,*Journal American statistics association* 65,1509-1526.

Chatfield, C. (2003). *The Analysis of Time Series: An Introduction,* 6th ed., New York, US: John Wiley and Sons.

Clements, M.P, and Hendry, D.F. (2004). *A companion to economic forecasting*. Blackwell Publishing, oxford.

Damsleth, E. (1979). Interpolating missing Values in Time Series. *Scand, J. Statist*., 7, 3339.

Durbin. (1960). The fitting of time series models. *Review of institute of international statistics*, 28, 233-244.

Dickay ,D., Bell, and Miller, R. (1986). Unit root in time series models: Tests and implications, *The American statistician*, 40, No.1,12-2.

Dickay D.A, and Fuller W.A(1979). Distribution of the estimates of autoregressive time series with unit root, *J. Amer statist. Assoc.,* 74, 427-431.

Gardner, G., Harvey, A. C., Phillips, G. D. A. (1980). An Algorithm for Exact Maximum Likelihood Estimation of Autoregressive-Moving Average Models by means of Kalman. *Applied Statistics.* 29, 311-322.

Gomez, I. A., Burton, D. M, and Love, H. A. (1995). Imputing Missing Natural Resource Inventory Data and the Bootstrap. *Natural Resource Modeling,* 9(4), 299-328.

Granger,and Newbold (1986)*: Forecasting Economic Time Series*, Academic Press, New York.

Ferreiro, O. (1987). Methodologies for the estimation of missing observations in time series. *Statistics and Probability Letters*, 5, 65-69.

Fung, D.S.C. (2006). *Methods for the Estimation of Missing Values in Time Series*. Edith Cowan University.

Frances, P.H (1998). *Time series models for business and economic forecasting*, Cambridge University Press.

Hamilton, J. D. (1994). *Time Series Analysis. New Jersey.* USA: Princeton University Press.

Harvey, A. C. (2001). *Forecasting, Structural Time Series Models and the Kalman Filter.*Cambridge, UK: Cambridge University Press.

Harvey, A. C. and Pierce, R.G. (1984). Estimating Missing Observations in Economi Time Series. *Journal of the American Statistical Association*, 79, 125-131.

Hinich, M.J.(1982). Testing For Gaussianity and linearity of stationary time series. *Time Series analysis*, 3, 169-176.

Janacek, G. and Swift, L. (1993). *Time Series Forecasting Simulation & Application,* West Sussex, England: Ellis Horwood Limited.

Jenkins, G.M. (1979). *Practical experiences with modeling and forecasting time series.* Gwilyn Jenkins and Partners Ltd, Jersey.

Jones, R. H. (1980). Maximum Likelihood Fitting of ARMA Models to Time Series with missing observations. *Technometrics*, 22, 389-395.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Energineering,* 81, 35-45.

 

Kihoro, J.M. (2006). *Imputation of missing data in seasonal Time series: comparative study under some parametric and non parametric methods*. Jkuat.

Kihoro J. M. (1998). *Estimation of missing observation in seasonal time series*. KenyattaUniversity.

Keenan, D,M.(1985). A turkey non additivty-type test for time series non-linearity. *Bometrika,* 72, 39-44

Kohn,and R Ansley, C. (1986). Estimation, prediction and interpolation for Arima models with missing data. *Journal American statistical Association*.

Little, and Rubin, D. (1987). *Statistical analysis with missing data.* Wiley and Sons New York.Luceno A. (1997). Estimation of missing values in possibly partially non stationary vectortime series. *Biometrika.*

Ljung, G.M. (1889). Estimation of missing values in time series. *Communication Statistics,*B 18, 459-465.

Makridakis, S. (1998), Forecasting: methods and applications, John Wiley and Sons, NewYork.

Mills, T.C. (1990): *Time series techniques for economists,* Cambridge University Press. Nieto, F. H., Martfncz, J. (1996). A Recursive Approach for Estimating Missing Observations in An univariate Time Series. *Communications in statistics Theory* A, 25, 2101-2116.

Newbold, P. and Granger, C.W.J (1979). Experience with forecasting univariete time Series and the combination with forecast.*J.R.A,*137,131-65.

Pena, D.and Tiao, G. C. (1991). A Note on Likelihood Estimation of Missing Values in Time Series. *The American statistician*, 45, 212-213.

Priestly,M.B.(1981). *Spectral analysis and time series*, Academic press, London.

Priestly,M.B (1988). *Non-linear & Non-stationary time series analysis*, London:Academic Press.

Rosen, Y. and Porat, B. (1989). Optimal ARMA Parameter Estimation Based on The Sample Covariances for data with Missing Observations. *IEEE Transactions on Information Theorey,* 35, 342-349.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data. Ann. Vol 72 of monographs on satatistics and applied Probability*. Chapman and Hall series.

Schwarz, G.(1978). Estimating the dimensional of model. *Ann. Statist*., 6,461-464

Shively, T. S. (1992). Testing for Autoregressive Disturbances in a Time Series Regression with Missing Observations. *Journal of econometrics*, 57, 233-255.

Subba Rao,T. and M.M.Gabr. (1980). A test for non-linearity of stationary time series.*Time Series Analysis*, 1, 145-158.

Shumway, R. H. (1982). An Approach to Time Series Smoothing and Forecasting using The EM Algorithm. *Journal of Time Series Analysis*, 3, 253-264

Tong, H.(1990).*Non linear time series. Dynamical System Approach*, Oxford University Press.

Tsay, R.S.(1986). Non linearity test for time series. *Biometrika,* 73,461-466.

Wei, W. (1989). *Time series analysis.* Wesley- publishing Company, New York

Worthke, W.(1998). Longitudinal and multi-group modeling with missing data in J.T.D. Little K.U. Schnabel, ed., *Modeling longitudinal and multi-group data*:*Practical issues*, applied approaches and specific examples, Mahwah,NJ: Lawrence Erlbaum Associates.