

Time Series Modeling and Forecasting of Somaliland Consumer Price Index: A Comparison of ARIMA and Regression with ARIMA Errors

Jama Mohamed

Faculty of Mathematics and Statistics, College of Applied and Natural Science, University of Hargeisa, Hargeisa, Somaliland

Email address:

jama.mohamed@live.co.uk

To cite this article:

Jama Mohamed. Time Series Modeling and Forecasting of Somaliland Consumer Price Index: A Comparison of ARIMA and Regression with ARIMA Errors. *American Journal of Theoretical and Applied Statistics*. Vol. 9, No. 4, 2020, pp. 143-153. doi: 10.11648/j.ajtas.20200904.18

Received: June 11, 2020; **Accepted:** June 22, 2020; **Published:** July 13, 2020

Abstract: In recent years, the Consumer Price Index (CPI) prediction has attracted the attention of many researchers due to its excellent measurement of macroeconomic performance. It is an important index that is used to measure the rate of inflation or deflation of commodities. In this paper, Autoregressive Integrated Moving Average (ARIMA) and regression with ARIMA errors, where the covariate is the time, were compared to forecast Somaliland Consumer Price Index using monthly time series data from 2013 – 2020. The study used and applied both models to produce the necessary forecasts. Also, Akaike Information Criterion (AIC), Corrected Akaike Information Criterion (AICc), Bayesian Information Criterion (BIC) and other model accuracy measures were used to measure model's predictive ability. By utilizing these methods, it is obtained that ARIMA (0, 1, 3) is the most suitable model for predicting CPI in Somaliland. Furthermore, the diagnostic tests show that the model presented is reliable and appropriate for forecasting Somaliland CPI data. The study results obviously indicate that CPI in Somaliland is more likely to proceed on an upward trend in the coming year. The study guides policymakers to use strict monetary and fiscal policy measures to address Somaliland's inflation.

Keywords: Augmented Dickey-Fuller Test, Autocorrelation, Autoregressive Integrated Moving Average (ARIMA), Consumer Price Index (CPI), Ljung-Box Test

1. Introduction

Consumer price index (CPI) is the most common economic indicator and measures the changes in prices of a group of goods over time. Therefore, it measures shifts in the purchasing power of money [1, 2]. Costa defined CPI as a weighted aggregate index which is calculated and published monthly [3]. In Somaliland, Ministry of National Planning and Development (MoNPD) publishes and compiles CPI every month in direct collaboration with Central Bank. Since 2013, the CPI has been the most often used measure of inflation in Somaliland. The Somaliland Annual Headline Inflation is estimated at 6.3 percent for the year ending April 2020 compared to the 5.6 percent reported for the year ending March 2020 [4]. In particular, there was a consistent monthly increase in prices of items beyond November 2016. The CPI was highest in April 2020. This is a good indication of higher inflation for the last three years in Somaliland.

The CPI is one of the most important variables for analyzing macroeconomic data in Somaliland. The main objective of the Somaliland monetary authorities like Central Bank is to fight inflation and maintain stable prices as inflation is measured directly from CPI data. The negative effects of inflation are well understood, which can contribute to a decline in the national currency's purchasing power leading to deteriorating socioeconomic conditions and living standards [5]. However, to recognize the factors that will determine its development in the near future, make the policymakers and domestic and foreign investors face a major challenge [6]. Such information would allow the Central Bank to predict future macroeconomic development and respond appropriately to economic shocks [6].

The CPI data is a time series data because it can be generally ordered depending on a time sequence. The advantage of this data type is that it can be forecasted. Time series data forecasting is a prediction by using the relationship pattern

between the variables based on time [7, 8]. CPI time series has an internal dynamic system that is regulated by itself, such that the time series fluctuation follows a particular order [9]. Autoregressive Integrated Moving Average (ARIMA) model is one of the methods widely used for predicting historical data [10-12]. Box and Jenkins developed this model and it is known as the Box-Jenkins time series method [13]. They strongly suggest that ARIMA process forecasts be made using the Difference Equation method, as this is the simplest approach.

A considerable number of studies were done to analyze the CPI data. For instance, Subhani and Panjawani studied on a monthly response to CPI announcements by government bonds [2]. Also, Adam, Awujola and Alungudu used the ARIMA model to study CPI in Nigeria [14]. Hamid and Dhakar studied seasonality analysis in the monthly US CPI data from January 1913 to December 2003 [15]. Similarly, Zhang, Che, Xu and Xu studied by ARMA model to analyze, model and forecast CPI time series data in China from January 1995 to May 2008 [16]. A research conducted by Kharimah, Usman, Widiarti and Elfaki further found that ARIMA (1, 1, 0), compared with ARIMA (0, 1, 1) and ARIMA (1, 1, 1), as the best and accurate model for forecasting CPI in Bandar Lampung, Indonesia [17]. In this research study, the univariate Box-Jenkins technique and regression with ARIMA errors will be used to analyze and predict the future value of Somaliland monthly CPI data.

2. Methodology

2.1. Data Description

The data used in this study is secondary data. This secondary data was obtained from the Central Statistics Department (CSD) of the MoNPD website (www.somalilandcsd.org). For the research purpose, monthly CPI data from January 2013 to April 2020 was used to forecast Somaliland CPI by using R software (version 3.6) especially forecast package. The statistical summaries, as well as time series distribution, would be tested using the skewness and kurtosis coefficients to check the presence of typical stylized data.

2.2. Box-Jenkins (ARIMA) Models

The Box-Jenkins (ARIMA) model is the most general class of time series prediction models in theory and was first popularized by Box and Jenkins [13]. $ARIMA(p, d, q)$ ignores independent variables and assumes the prior series values plus previous error terms provide information for forecasting purposes. The integers apply to the data set parts of the Autoregressive (AR), Integrated (I), and Moving Average (MA), respectively. In certain cases, the models are applied to data showing signs of non-stationarity which can be stationarized by transformation such as detrending and logging. The model takes historical data into account and breaks it down into AR process, where there is a memory of past events; an integrated process that accounts for stationarity, making it easier to predict; and an MA of forecast errors, so that the longer the historical data, the more reliable the predictions will

be because they learn over time. The ARIMA models only apply to a stationary data series where the function of mean, variance, and autocorrelation remains constant over time.

AR process expresses a response variable as a function of the response variable's past values. A p^{th} -order AR process is given by:

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t \quad (1)$$

Where y_t is the stationary response variable being forecasted at time t , $y_{t-1}, y_{t-2}, \dots, y_{t-p}$ is the dependent variable at time lags $t-1, t-2, \dots, t-p$, $\alpha = \mu(1 - \phi_1 - \phi_2 - \dots - \phi_p)$, $\phi_1, \phi_2, \dots, \phi_p$ are the parameters to be estimated and ε_t is the error term at time t with mean zero and a constant variance. Using the backshift operator, we can define the $AR(p)$ process as:

$$(1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p) y_t = \varepsilon_t \quad (2)$$

The MA process of order q , $MA(q)$, can be written in the form:

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q} \quad (3)$$

Where q is the number of lags in the moving average and $\theta_1, \theta_2, \dots, \theta_q$ are the coefficients to be estimated. The backshift operator of MA process can be represented as:

$$\theta(B) = \theta_1 B + \theta_2 B^2 + \dots + \theta_q B^q \quad (4)$$

To construct an ARMA model, we start with a regression equation with no explanatory variables, $y_t = \beta_0 + \varepsilon_t$ and add AR process as well as MA process to it.

$$y_t = \beta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q} \quad (5)$$

Where $\beta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p}$ is an $AR(p)$, $\varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q}$ is a $MA(q)$, ϕ_s and θ_s are the parameters of the AR and MA processes respectively.

The ARIMA or integrated ARMA model is an extended category of ARMA model that contains a differencing term. If

$$\nabla^d y_t = (1 - B)^d y_t \quad (6)$$

is an $ARMA(p, q)$, then the process is said to be $ARIMA(p, d, q)$. This is generally written as:

$$\phi(B)(1 - B)^d y_t = \theta(1 - B)\varepsilon_t \quad (7)$$

A first-differenced CPI series is of the form:

$$CPI_t = (\nabla CPI_t) = CPI_t - CPI_{t-1} = \Delta CPI_t - \Delta CPI_{t-1} \quad (8)$$

Therefore the $ARIMA(p, 1, q)$ model may be stated as:

$$CPI_t = \beta_0 + \phi_1 CPI_{t-1} + \phi_2 CPI_{t-2} + \dots + \phi_p CPI_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q} \quad (9)$$

Where CPI_t is the differenced CPI series of first order, and ϕ, β and θ are the coefficients to be estimated. Before applying to a time series model, the equation has to assume stationarity. Successive differences are taken in case of non-stationarity before the sequence is stationary. In practice, the differences are rarely greater than two.

The Box-Jenkins time series approach seeks to identify the most suitable $ARIMA(p, d, q)$ model and use it for forecasting purposes. It utilizes a six-phase iterative scheme:

A priori determination of the order of differentiation, d (or selection of appropriate transformation);

A priori determination of the orders of AR and MA processes, p and q ;

Model identification;

Estimation of the model parameters (ϕ, β and θ , $\sigma^2 = \text{var}(\varepsilon_t)$);

Diagnostic checking (validation);

Forecasting

2.3. Testing Autocorrelation

One of the assumptions of time series regression is that the errors are independent. Error terms correlated over time are said to be serially correlated or autocorrelated. Serial correlation of the disturbances with Ordinary Least Square (OLS) estimation can have the following effects:

Estimated coefficients of regression are still unbiased but no longer have the minimum variance property (inefficient);

The OLS estimate of s^2 (MSE) could underestimate the true error variance;

The true standard error of estimate could be underestimated by $se\{b_k\}$;

Statistical inferences using t and F tests are also no longer valid.

So, it is important to test the existence of serial correlation. In general, there are two ways of detecting autocorrelation. The first is the informal way which is done through graphs (plotting residuals against time) and therefore we call it the graphical method. The second is through formal tests for autocorrelation, like Durbin Watson Test. The Durbin Watson Test is used to test the hypothesis: $H_0: \rho = 0$ against $H_1: \rho \neq 0$. The Durbin Watson statistic is:

$$D = \frac{\sum_{t=2}^T (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^T \varepsilon_t^2} \quad (10)$$

The value of Durbin Watson test statistic is compared with the critical value relevant to it. If the test statistic is less than critical value then we reject the null hypothesis and conclude that there is autocorrelation, i.e. If $D < d_L$, reject $H_0: \rho = 0$

and accept $H_1: \rho \neq 0$; If $D > d_U$, do not reject $H_0: \rho = 0$ and if $d_L < D < d_U$, the test is inconclusive.

2.4. Testing Stationarity

To model a time series data, we examine the data structure to obtain some preliminary information about the stationarity of the series; whether a trend or seasonal pattern exists. If both the mean and variance are constant over time, a time series is said to be stationary. A time series plot of the data is suggested to decide whether any differencing is required before formal tests are conducted. If the data is non-stationary, we perform a Box Cox transformation or take the series' first (or higher) order difference which may result in a stationary time series. Data differentiation times are indicated by parameter d in the model, $ARIMA(p, d, q)$. An Augmented Dickey-Fuller (ADF) test is then used to evaluate the stationarity of the results.

2.4.1. Augmented Dickey-Fuller Test

The ADF test for unit root is powerful test used to check whether a time series is stationary or not. The ADF test procedure is similar to the Dickey-Fuller test except it is implemented to the model. A random walk with trend and drift is defined as follows:

$$\Delta y_t = \alpha + \beta_t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \dots + \delta_p \Delta y_{t-p+1} + \varepsilon_t \quad (11)$$

Where α is a constant, β is the parameter on a time trend and p is the lag order of the AR process. Substituting $\alpha = 0, \beta = 0$ into eq. (11) corresponds to random walk model and using $\beta = 0$ corresponds to random walk model with drift.

The test statistic, τ value is given by:

$$\tau = \frac{\hat{\gamma}}{\hat{\sigma}_{\hat{\gamma}}} \quad (12)$$

Where $\hat{\gamma}$ and $\hat{\sigma}_{\hat{\gamma}}$ are the coefficient estimate and the standard error of the estimated coefficient respectively. The null hypothesis for an ADF test is: $H_0: \gamma = 0$ versus the alternative hypothesis $H_1: \gamma < 0$. To put it another way, the null hypothesis is that the data has a unit root while the alternative hypothesis is that the data does not have a unit root. The value of the test statistic is compared with the critical value relevant for the Dickey-Fuller test. If the test statistic is smaller than the critical value then we reject the null hypothesis and deduce that there is no unit root. The ADF test fails to test for stationarity explicitly, but indirectly through the presence (or absence) of a unit root. Using the normal threshold of 5%, differencing is needed if the p-value exceeds 0.05.

2.4.2. Correlograms

In addition to graphical stationarity checking, formal test schemes are implemented using autocorrelation function (ACF) and partial autocorrelation function (PACF). The correlograms examine the data from the time series by plotting the ACF and PACF to try and obtain the data's functional form.

The ACF reflects the degree of continuity over the respective variables lags; a correlation at time y_t and y_{t+k} between two values of the same variable. The sample ACF can be written mathematically as follows:

$$r_k = \frac{\sum_{t=1}^{T-k} (y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^T (y_t - \bar{y})^2} \quad (13)$$

The PACF calculates the degree of association of two variables that is not described with given set of other variables by their mutual correlations. The sample PACF is:

$$p_{k,k} = \begin{cases} r_1 & \text{if } k=1, \\ r_k - \frac{\sum_{j=1}^{k-1} (p_{k-1,j} \cdot r_{k-j})}{1 - \sum_{j=1}^{k-1} (p_{k-1,j} \cdot r_j)} & \text{if } k=2,3,\dots \end{cases} \quad (14)$$

Where $p_{k,j} = p_{k-1,j} - p_{k,k} p_{k-1,k-j}$ for $j=1,2,\dots,k-1$. The ACF will be used to decide the order of MA process while the PACF will determine the order of AR process. The key defining characteristics of the theoretical ACFs and PACFs for stationary processes are given in Table 1.

Table 1. Defining characteristics of ACF and PACF for stationary processes.

Process	ACF	PACF
AR	Decays exponentially like damped sine wave or tails off towards zero	Cuts off to zero after lag p
MA	Cuts off to zero after lag q	Decays exponentially like damped sine wave or tails off towards zero

If the original or differenced time series turns out to be non-stationary some suitable transformations will be made to achieve stationarity, then we must proceed to the next step where initial values are identified.

2.5. Model Identification and Estimation

Box-Jenkins approach is applied by observing the ACF and PACF of the time series. ACF and PACF are therefore at the heart of how to classify the ARIMA model. There are three rules for the identification of $ARIMA(p, d, q)$ model:

If ACF graph is cut off after lag q and PACF dies down, we recognize $MA(q)$ resulting in the model of $ARIMA(0, d, q)$.

If ACF graph dies down and PACF is cut off after lag p , $AR(p)$ resulting in $ARIMA(p, d, 0)$ is identified.

If ACF and PACF die down; the ARIMA model is mixed, differencing is necessary.

If the model order has been specified, (i.e., p , d and q values), the parameters ϕ , β and θ need to be estimated. In fitting the ARIMA model, the concept of parsimony is

important whereby the model should have the smallest possible parameters and still be able to explain the sequence (p and q should be 2 or less). The more parameters, the greater the noise that can be inserted into the model, and hence the greater the variance. Moreover, the following methods are also applied: maximum likelihood estimation (MLE), Akaike Information Criterion (AIC), Akaike Information Criterion Corrected (AICc) and Bayesian Information Criterion (BIC).

2.5.1. Maximum Likelihood Estimation (MLE)

The ARIMA model will be estimated using maximum likelihood estimation (MLE) technique. This method obtains parameter values which maximize the likelihood of obtaining the data we observed. The MLE is quite similar to the least-square estimation for ARIMA models. The likelihood function in a standard Gaussian is:

$$\text{Log}L = -\frac{T}{L} \ln(2\pi) - \frac{T}{L} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^T \varepsilon_i^2 \quad (15)$$

Where T is the time $t = 1, 2, \dots, T$ of the time series data, σ and ε are the constant variance and the error terms respectively. The log likelihood presents the logarithm of the probability of the observed data from the fitted model. We select the model with maximum log likelihood.

2.5.2. Information Criteria

The AIC is useful in deciding the order of an ARIMA model. It is used for the comparison of competing models that fit the same series. It can be defined as follows:

$$AIC = -2 \ln L + 2p \quad (16)$$

Where L is the likelihood of the data and p is the number of fitted model parameters (including the residual variance). The original representation of AIC applies a linear penalty term to the number of free parameters, but the AICc introduces a second term to factor into the sample size, making it more appropriate for smaller sample sizes. The AICc (corrected for small sample bias) is given by:

$$AICc = AIC + \frac{2p(p+1)}{n-p-1} \quad (17)$$

The Bayesian Information Criterion (BIC) can be defined as:

$$BIC = AIC + p(\ln(n) - 2) \quad (18)$$

In general, the BIC penalizes free parameters stronger than the AIC, although it depends on the size of n and relative magnitude of n and p . Potential models are obtained by decreasing the AIC, AICc or BIC and optimizing log likelihood. Our choice is to use the AICc and to choose the parsimonious model with the smallest AICc and the greatest log likelihood.

2.6. Model Validation and Forecasting

Estimated model(s) would be considered the most suitable

if it usually simulates historical behavior and constitutes white-noise innovations. Historical behavior will be checked by the ACF and PACF of estimated series and pick the one that better describes the temporal dependency in the CPI series, i.e. the model(s) whose residuals do not display significant lags. White noise innovations will be checked as well as overfitting through a series of diagnostic tests based on projected residuals. The Ljung-Box test can also be used to verify if a time series autocorrelation varies from zero. If the result rejects the hypothesis, that implies that the data is independent and uncorrelated; otherwise, serial correlation persists in the sequence and the model needs adjustment.

A good feature of the ARIMA class is its power of forecasting. Gujarati believed that ARIMA's popularity was due to its prediction success [18]. To forecast future values of the time series, we use this equation:

$$\hat{y}_{T+h} = E[y_{T+h} | y_T, y_{T-1}, \dots, \varepsilon_t, \varepsilon_{t-1}, \dots] = \sum_{j=1}^p \phi_j \hat{y}_{T+h-j} + \sum_{j=1}^q \theta_j \hat{\varepsilon}_{T+h-j} \quad (19)$$

Where t is the past until T , \hat{y}_{T+h} is the forecast of y_{T+h} and $\hat{y}_T(h)$ is found from the equation of the model by substituting the future value of ε_t into zero, the future value of y into their conditional mean and past values of y and ε_t into their actual values. The standard error of the forecast error is given by:

$$\sigma_T(h) = \sqrt{\sigma_\varepsilon^2 \sum_{j=0}^{h-1} \phi_j^2} \quad (20)$$

Thus, supposing that forecast errors are normally distributed, a $(1-\alpha)100\%$ prediction interval for the future values, y_{T+h} , can be created as follows:

$$\hat{y}_T(h) \pm z_{\alpha/2} \sigma_T(h) \quad (21)$$

2.6.1. The Ljung-Box Test

Ljung and Box developed the standard portmanteau test to test that the data is a realization of a powerful white noise [19]. It is about calculating the following statistic

$$Q(m) = n(n+2) \sum_{k=1}^m \frac{r_k^2}{(n-k)} \quad (22)$$

and rejecting the powerful white noise hypothesis if $Q(m)$ is larger than $(1-\alpha)$ quantile. n is the sample size, r_k is the autocorrelation of the sample at lag k and m is the lag order that should be stated. This is a one-tail (i.e. one-sided) test, so that the calculated p-value should be compared to the entire significance level (α). In practice, the selection of m will affect the output of the $Q(m)$ statistic. A few values of m are commonly used. Simulation studies suggest that better power

performance is provided by choosing $m \approx \ln(T)$. The null hypothesis is H_0 : The data are distributed independently (i.e. the correlations in the population from which the sample is taken are 0, and any correlations found in the data result from the sampling process being random). The alternative hypothesis is H_1 : Data are not distributed independently. If the model's Ljung-Box statistic is not significantly different from zero then we do not reject the null hypothesis of no remaining significant autocorrelation in the model's residuals and conclude that the model appears adequate to capture the time series correlation information. i.e. if $Q \leq \chi^2_{\alpha}(m-p-q)$, then H_0 is not rejected; if $Q > \chi^2_{\alpha}(m-p-q)$, then H_0 is rejected. The best fitting model(s) will then go through different residual and normality checks and only suitable model(s) will be chosen for the purpose of forecasting.

2.6.2. Forecasting Accuracy Measures

The forecasting performance of the different types of ARIMA models is compared by calculating forecasting accuracy statistics like Mean Error (ME), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) and Theil's U. The lower the figures, the better the model and therefore, the predicted values are deemed accurate.

$$ME = \frac{1}{H} \sum_{h=1}^H (CPI_{T+h} - \hat{CPI}_T(h)) \quad (23)$$

$$MAE = \frac{1}{H} \sum_{h=1}^H |CPI_{T+h} - \hat{CPI}_T(h)| \quad (24)$$

$$RMSE = \sqrt{\frac{1}{H} \sum_{h=1}^H (CPI_{T+h} - \hat{CPI}_T(h))^2} \quad (25)$$

$$MAPE = \frac{1}{H} \sum_{h=1}^H \left| \frac{CPI_{T+h} - \hat{CPI}_T(h)}{CPI_{T+h}} \right| * 100 \quad (26)$$

$$U = \frac{\sqrt{\frac{1}{H} \sum_{h=1}^H (CPI_{T+h} - \hat{CPI}_T(h))^2}}{\sqrt{\frac{1}{H} \sum_{h=1}^H CPI_{T+h}^2} + \sqrt{\frac{1}{H} \sum_{h=1}^H \hat{CPI}_T^2(h)}} \quad (27)$$

3. Results and Discussion

The Somaliland CPI data was subjected to descriptive statistics. Table 2 shows the descriptive statistics of the CPI series. The total number of observations was 88, the highest reported CPI was 183.56 recorded in April 2020 and the lowest reported CPI was 106.27 recorded in January 2013. The skewness is 0.29 implying that the CPI series is negatively skewed and non-symmetric. The estimated kurtosis was obtained to be -1.44 indicating that the CPI series is not

normally distributed.

Table 2. Descriptive statistics of Somaliland CPI data.

Statistic	Value
No. of observations	88
Minimum	106.27
1 st Quartile	114.36
Median	131.73
Mean	139.09
3 rd Quartile	166.42
Maximum	183.56
Range	77.29
Standard deviation	25.42
Skewness	0.29
Kurtosis	-1.44

Figure 1 shows a time series plot of Somaliland CPI series from January 2013 to April 2020. The figure clearly exhibits an upward trend in the monthly CPI. To discuss it in more detail, the CPI gradually increased in the years 2013 up to 2016. In April 2016, it decreased until October 2016 and after that it began to rise sharply. In the last three years, CPI has increased suddenly up to 40 percent higher than the previous years. For this reason, the mean definitely does not seem to be constant and therefore the series is not stationary. In addition, there is no seasonal variation from the series and thus there is no seasonal part from the data.

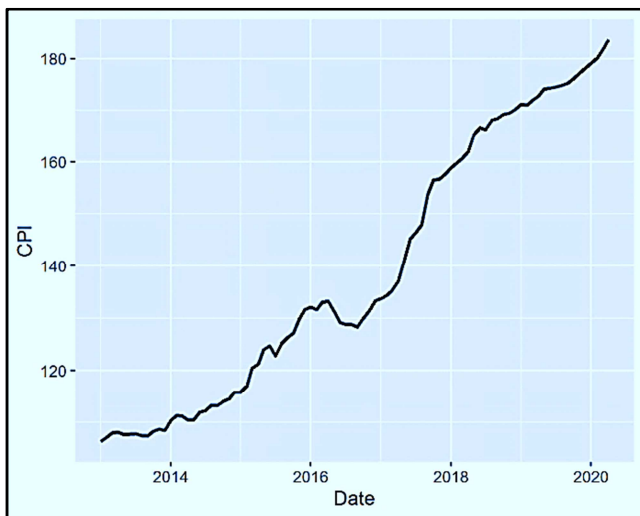


Figure 1. Time series plot of Somaliland CPI (January 2013 – April 2020).

Before we test stationarity, an important step is to test the independence of errors assumption. From table 3, we observe that the Durbin Watson statistic is 0.0580 with a p-value of approximately zero. We conclude that the CPI data has positive autocorrelation.

Table 3. Testing autocorrelation using Durbin-Watson test.

Durbin-Watson test
DW = 0.058047, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0

Next, we proceed to test stationarity in the formal way. It can be checked by the absence or presence of unit root. Table 4 shows the unit root test for determining the stationarity of the

series. We see that the Dickey-Fuller statistic is -2.2425 with a p-value of 0.4764. Then, we assume that the CPI series is non-stationary time series.

Table 4. Augmented Dickey-Fuller Test for unit root of CPI series.

Augmented Dickey-Fuller Test
data: CPI\$CPI
Dickey-Fuller = -2.2425, Lag order = 1,
p-value = 0.4764
alternative hypothesis: stationary

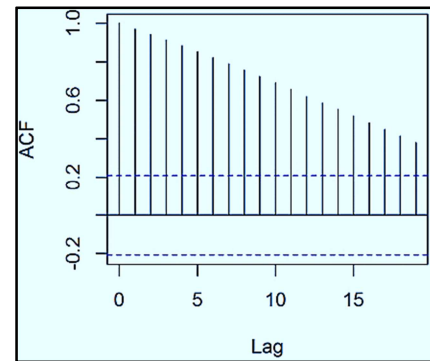


Figure 2. Autocorrelation function plot of the CPI series.

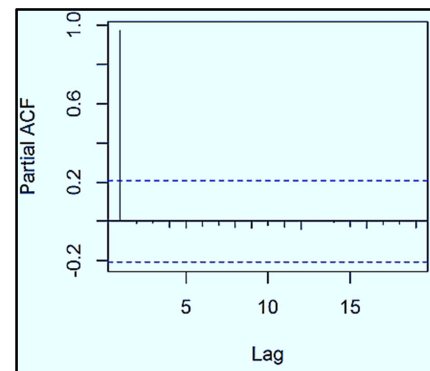


Figure 3. Partial autocorrelation function plot of the CPI series.

Figures 2 and 3 show ACF and PACF of the CPI series. The coefficients of ACF start with a high value and declines slowly as lags increase, indicating that the series is non-stationary. The spikes in the ACF plot that cross above the cut-off line suggest that the current level of CPI is significantly autocorrelated with its lagged values. The corresponding PACF plot only has a significant spike at lag 1 and then cuts off, which means that the autocorrelations at lag 2 and above are solely due to the outbreak of autocorrelation at lag 1. The non-stationarity is of the order one since only the first-lagged bar is considerably higher than the critical limit i.e. the first lag of PACF plot is above the significant line. This implies non-stationarity and suggests differencing of the first order as the remedy.

Since the Somaliland CPI data is non-stationary, it must be configured in first differences to become stationary. The series was transformed by taking the first differences of the values in the series so as to attain stationarity in the first moment. The equation representing the transformation is given by:

$$\Delta CPI_t = CPI_t - CPI_{t-1} \quad (28)$$

where CPI_t represents the monthly values for the CPI series.

The time series plot for the differences of the series is presented in Figure 4. The mean appears to be constant over time, i.e. there is no trend increment.

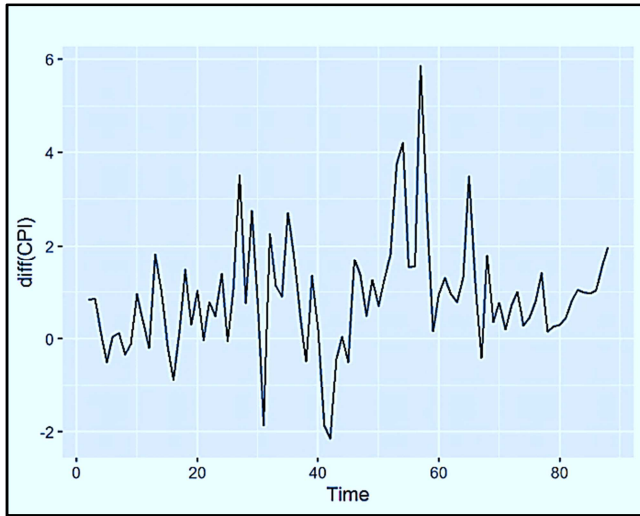


Figure 4. Time series plot for the first difference of the CPI series.

Table 5 shows the unit root test for determining the stationarity of the first differences. Again, we observe that the Dickey-Fuller statistic is -6.3802 with a p-value of less than 0.01, so we reject the null hypothesis that the first difference series is not stationary. This means we assume that the CPI first differences are stationary time series.

Table 5. Augmented Dickey-Fuller test for unit root of CPI first difference.

Augmented Dickey-Fuller Test
data: diff
Dickey-Fuller = -6.3802, Lag order = 0,
p-value = 0.01
alternative hypothesis: stationary
Warning message:
In adf.test (diff, k = 0): p-value smaller than printed p-value

Figures 5 and 6 show the ACF and PACF for the first difference of the CPI series. The spikes at lag 0, 1, 3, 13, 15 and 16 are beyond the significant line, so we can say that there is autocorrelation at lag 0, 1, 3, 13, 15 and 16. The bar at lag 5 and 6 is around zero, so the CPI first difference is stationary.

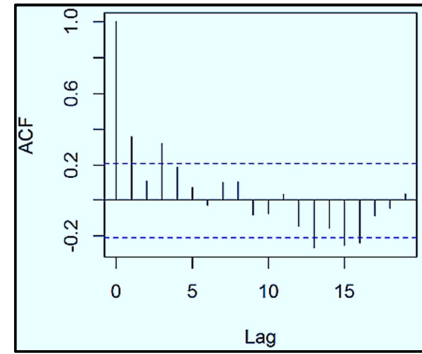


Figure 5. Autocorrelation function plot for the first difference of the CPI series.

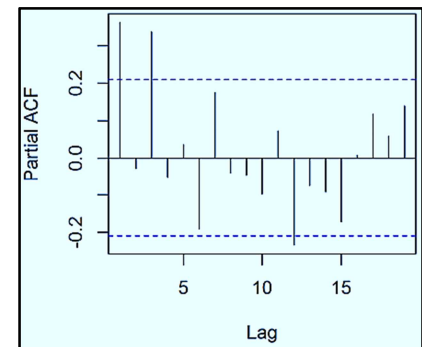


Figure 6. Partial autocorrelation function plot for the first difference of the CPI series.

3.1. Utilizing Box-Jenkins Methodology

We assumed that the differenced CPI series is stationary, so it is safe to identify an ARIMA model and estimate its parameters for the CPI series.

Table 6 shows the results obtained from ARIMA models with different orders. A model with the greatest log-likelihood and lowest AIC, AICc and BIC is better than the other models. Theil's U lies 0 to 1, the nearer it gets to zero, the better the forecast technique [20]. The study will consider the log-likelihood, AIC, AICc and BIC, Theil's U, ME, MAE, RMSE and MAPE only as the criteria for forecasting CPI in Somaliland, and thus the ARIMA (0, 1, 3) is carefully chosen. This model is the only one to meet the above-mentioned conditions and the parsimony principle which prioritizes the smallest parameter possible in the model.

Table 6. Results from different ARIMA models.

Model	Log-likelihood	AIC	AICc	BIC	U	ME	MAE	RMSE	MAPE
ARIMA (2, 1, 2)	-128.14	268.28	269.33	283.07	0.0074	0.0059	0.7672	1.0440	0.5780
ARIMA (0, 1, 0)	-140.74	285.47	285.62	290.40	0.0086	0.0012	0.8393	1.2129	0.6297
ARIMA (1, 1, 0)	-134.67	275.34	275.63	282.74	0.0080	0.0015	0.8006	1.1303	0.6027
ARIMA (0, 1, 1)	-133.15	272.29	272.58	279.69	0.0078	0.0016	0.8256	1.1095	0.6242
ARIMA (1, 1, 2)	-130.14	270.28	271.02	282.61	0.0076	0.0011	0.7810	1.0688	0.5907
ARIMA (2, 1, 1)	-129.84	269.68	270.42	282.00	0.0075	0.0011	0.7819	1.0649	0.5910
ARIMA (3, 1, 2)	-127.43	268.85	270.27	286.11	0.0073	0.0051	0.7586	1.0354	0.5701
ARIMA (0, 1, 3)	-127.73	265.46	266.20	277.79	0.0073	0.0008	0.7390	1.0387	0.5564
ARIMA (0, 1, 2)	-131.96	271.92	272.41	281.79	0.0077	0.0015	0.8185	1.0928	0.6187
ARIMA (0, 1, 4)	-127.03	266.07	267.12	280.86	0.0073	0.0010	0.7462	1.0304	0.5611

After selecting the model, the parameters or coefficients of the model need to be estimated. As shown in Table 7, The

coefficients of the MA (1), MA (3) and drift components are positive and statistically significant at 1% level of significance,

while the coefficient of the MA (2) component is negative and not statistically significant at 1%, 5% and 10% levels of significance. This indicates that undetected CPI shocks have a positive impact on current CPI in Somaliland. Such shocks can include but not limited to, shocks from monetary policy

and favorable political outcomes. In fact, the results show that an increase of 1 percent in these shocks would lead to an increase in CPI of around 0.42%, 0.34% and 0.90% respectively, hence higher inflation.

$$\Delta CPI_t = 0.0418304\varepsilon_{t-1} - 0.055791\varepsilon_{t-2} + 0.335884\varepsilon_{t-3} + 0.897600 \quad (29)$$

Table 7. Estimation of ARIMA (0, 1, 3) parameters.

ARIMA (0, 1, 3) Model:				
Variable	Coefficient	Standard Error	z value	Pr (> z)
MA (1)	0.418304	0.109215	0.188582	0.0001281***
MA (2)	-0.055791	0.097564	-0.5718	0.5674302
MA (3)	0.335884	0.112359	2.9894	0.0027955***
Drift	0.897600	0.188582	4.7597	1.938e-06***

P-value: (0.0001281) (0.5674302) (0.0027955) (1.938e-06)

S.E: (0.109215) (0.097564) (0.112359) (0.188582)

Before model validation, we carefully look at the stability of the chosen model. In Figure 7, the inverse roots of MA characteristic polynomials for the stability of the ARIMA (0, 1, 3) model are presented. As conventionally expected, we see that the ARIMA model is stable as the accompanying inverse roots of the characteristic polynomials are in the unit circle. This illustrates that our model is reliable and most appropriate for forecasting CPI in Somaliland over the period under study.

As in Figure 8, the ACF plot of the residuals from the ARIMA (0, 1, 3) model shows that all autocorrelations are within threshold levels, indicating the residuals act as white noise. In table 8, A Ljung-Box test returns a great p-value, also suggesting that the residuals are white noise.

Table 8. Results from Ljung-Box test of residuals from ARIMA (0, 1, 3) with drift.

Ljung-Box test
data: Residuals from ARIMA (0, 1, 3) with drift
Q* = 5.4193, df = 6, p-value = 0.4913
Model df: 4. Total lags used: 10

As the purpose of this study, after identifying, estimating and validating the model, it is important to forecast the future values. Table 9 and Figure 9 (with a projected range from May 2020 to April 2021) simultaneously show that CPI is expected to continue growing sharply in Somaliland over the next year.

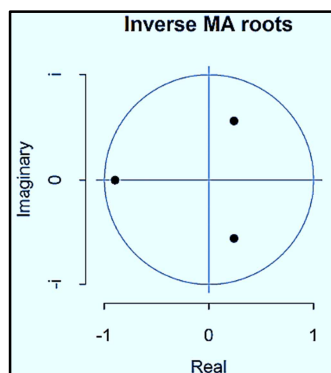


Figure 7. Inverse roots of MA characteristic polynomial (s).

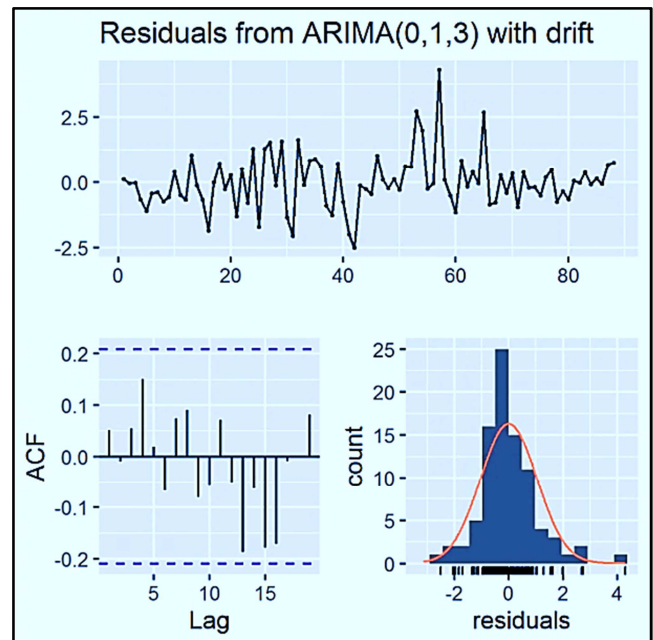


Figure 8. Diagnostic checking of residuals from ARIMA (0, 1, 3) with drift.

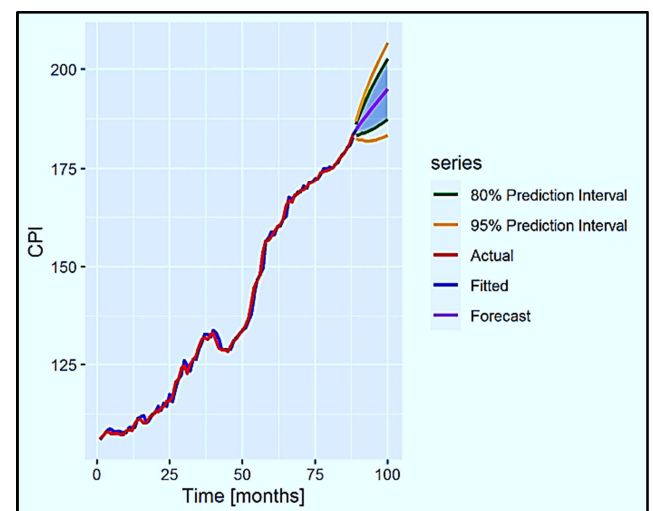


Figure 9. Predicted monthly CPI values from ARIMA (0, 1, 3).

Table 9. Future forecasts from ARIMA (0, 1, 3).

Date	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
May 2020	184.7128	183.3421	186.0835	182.6165	186.8091
Jun 2020	185.7949	183.4162	188.1735	182.1570	189.4327
Jul 2020	186.9449	183.9206	189.9691	182.3197	191.5700
Aug 2020	187.8425	184.0260	191.6589	182.0057	193.6792
Sep 2020	188.7401	184.2696	193.2105	181.9031	195.5770
Oct 2020	189.6377	184.2696	194.6779	181.9293	197.3460
Nov 2020	190.5353	184.9834	196.0871	182.0444	199.0261
Dec 2020	191.4329	185.4127	197.4530	182.2258	200.6399
Jan 2021	192.3305	185.8758	198.7851	182.4590	202.2019
Feb 2021	193.228	186.3665	200.0896	182.7341	203.7220
Mar 2021	194.1257	186.8799	201.3714	183.0442	205.2071
Apr 2021	195.0233	187.4127	202.6338	183.3839	206.6626

3.2. Utilizing Regression with ARIMA Errors

In this section, we model the CPI data using regression with ARIMA errors as it can be a potential model for autocorrelated time series data.

Table 10 presents the output obtained from different

regression with ARIMA errors models. As mentioned before, a model with the lowest AIC, AICc, BIC, Theil's U, ME, MAE, RMSE and MAPE and the greatest log-likelihood will be selected. The priority of the selection criteria will be given to these forecasting accuracy measures and thus regression with ARIMA (2, 0, 3) errors is carefully selected.

Table 10. Results from different regression with ARIMA errors models.

ARIMA Errors	Log-likelihood	AIC	AICc	BIC	U	ME	MAE	RMSE	MAPE
ARIMA (2, 0, 2)	-132.25	278.50	279.90	295.84	0.0075	-0.0660	0.7849	1.0627	0.5970
ARIMA (1, 0, 0)	-143.10	294.21	294.69	304.12	0.0086	-0.0758	0.8352	1.2097	0.6306
ARIMA (1, 0, 2)	-134.06	280.13	281.17	294.99	0.0077	-0.0672	0.8230	1.0865	0.6253
ARIMA (2, 0, 1)	-132.67	277.34	278.38	292.21	0.0076	-0.0666	0.7918	1.0681	0.6015
ARIMA (1, 0, 1)	-135.06	280.12	280.85	292.51	0.0078	-0.0670	0.8272	1.1006	0.6293
ARIMA (2, 0, 0)	-136.47	282.95	283.68	295.34	0.0079	-0.0675	0.8048	1.1200	0.6091
ARIMA (3, 0, 1)	-131.81	277.63	279.03	294.97	0.0075	-0.0648	0.7832	1.0572	0.5951
ARIMA (3, 0, 0)	-136.47	284.95	285.99	299.81	0.0079	-0.0675	0.8048	1.1200	0.6092
ARIMA (2, 0, 3)	-128.16	272.32	274.14	292.13	0.0072	-0.0560	0.7647	1.0131	0.5772
ARIMA (2, 0, 4)	-128.48	274.96	277.26	297.25	0.0072	-0.0595	0.7664	1.0169	0.5792

After the model selection, we estimate the parameters of the chosen model. Table 11 shows that the coefficients of the Time, AR (1), AR (2), MA (3) and Intercept components are statistically significant at 5% level of significance, while the

coefficient of the MA (1) and MA (2) components are negative and not statistically significant at 1%, 5% and 10% levels of significance. Actually, the results show that an increase of 1 unit in time will increase the CPI by about 0.94.

$$CPI_t = 98.4250 + 0.9418time + 1.4576\eta_{t-1} - 0.5032\eta_{t-2} - 0.0325\epsilon_{t-1} - 0.2644\epsilon_{t-2} + 0.4266\epsilon_{t-3} \quad (30)$$

Table 11. Estimation of regression with ARIMA (2, 0, 3) errors parameters.

Regression with ARIMA (2, 0, 3) Errors Model:				
Variable	Coefficient	Standard Error	z value	Pr (> z)
Time	0.941839	0.078270	12.0332	< 2.2e-16***
Intercept	98.425049	4.383430	22.4539	< 2.2e-16***
AR (1)	1.457589	0.254189	5.7343	9.793e-09***
AR (2)	-0.503237	0.246423	-2.0422	0.04114*
MA (1)	-0.032473	0.238124	-0.1364	0.89153
MA (2)	-0.264438	0.163932	-1.6131	0.10672
MA (3)	0.426589	0.101447	4.2050	2.610e-05***

$$P - \text{value} : (< 2.2e - 16) (< 2.2e - 16) (9.793e - 09) (0.04114) (0.89153) (0.10672) (2.610e - 05)$$

$$S.E : (4.383430) (0.078270) (0.254189) (0.246423) (0.238124) (0.163932) (0.101447)$$

The next step is to look at the stability of the selected model. Figure 10 represents the inverse roots of AR and MA characteristic polynomials for the stability of regression with ARIMA (2, 0, 3) errors model. We can observe that the regression with ARIMA (2, 0, 3) errors model is stable since

the corresponding inverse roots of the characteristic polynomials are in the unit circle. This demonstrates that our model is stable and suitable for predicting CPI in Somaliland over the period under consideration.

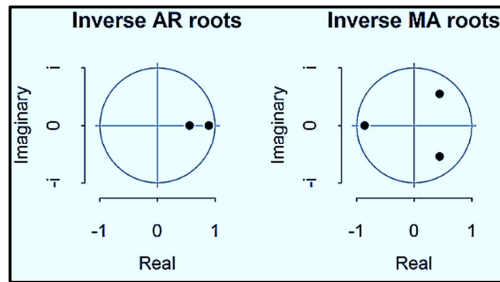


Figure 10. Inverse roots of AR and MA characteristic polynomial (s).

Before forecasting future values, we should validate the selected potential model. As in Figure 11, the ACF plot of the residuals from the regression with ARIMA (2, 0, 3) errors model shows that all autocorrelations are within the dashed lines, suggesting that the residuals behave like white noise. Also, in Table 12, a Ljung-Box test returns a large p-value, suggesting that the residuals are not significantly different from white noise.

Table 12. Results from Ljung-Box test of residuals from regression with ARIMA (2, 0, 3) errors.

Ljung-Box test
data: Residuals from Regression with ARIMA (2, 0, 3) errors
 $Q^* = 2.5646$, $df = 3$, $p\text{-value} = 0.4637$
Model df : 7. Total lags used: 10

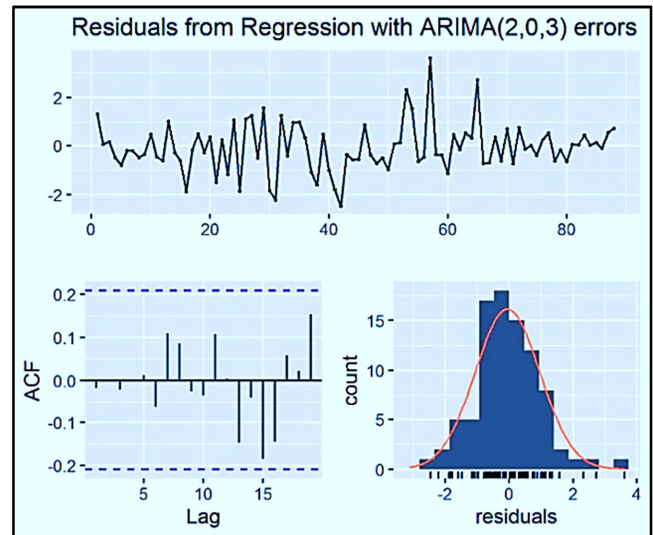


Figure 11. Diagnostic checking of residuals from regression with ARIMA (2, 0, 3) errors.

The last step is to forecast Somaliland monthly CPI in the short term (12 months ahead). As shown clearly in Table 13 and Figure 12 (with a projected range from May 2020 to April 2021), the CPI is likely to continue rising sharply in Somaliland over the next twelve months.

Table 13. Future forecasts from regression with ARIMA (2, 0, 3) errors.

Date	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
May 2020	184.6999	183.3467	186.0532	182.6303	186.7696
Jun 2020	185.6811	183.3252	188.0371	182.0780	189.2843
Jul 2020	186.8460	183.8979	189.7941	182.3373	191.3547
Aug 2020	187.7762	184.1038	191.4487	182.1597	193.3927
Sept 2020	188.5889	184.2556	192.9222	181.9617	195.2161
Oct 2020	189.3483	184.4654	194.2311	181.8806	196.8160
Nov 2020	190.0892	184.7648	195.4136	181.9462	198.2322
Dec 2020	190.8300	185.1557	196.5043	182.1519	199.5081
Jan 2021	191.5799	185.6296	197.5302	182.4797	200.6801
Feb 2021	192.3431	186.1753	198.5109	182.9103	201.7760
Mar 2021	193.1212	186.7818	199.4606	183.4259	202.8165
Apr 2021	193.9143	187.4392	200.3894	184.0116	203.8171

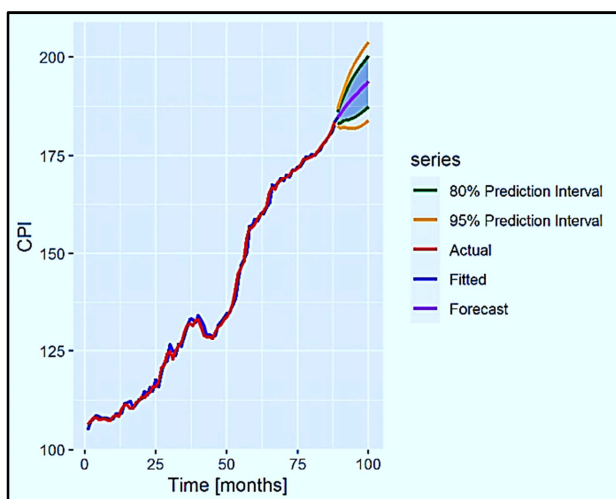


Figure 12. Predicted monthly CPI values from regression with ARIMA (2, 0, 3) errors.

3.3. Comparative Analysis

In this comparative study, it is obvious that ARIMA (0, 1, 3) and regression with ARIMA (2, 0, 3) errors models are competing with each other. However, we strongly recommend ARIMA (0, 1, 3) as the ideal model since it captures the stochastic variation in the data better than the other model, i.e. it has the smaller information criteria, prediction error and model parameters.

4. Conclusion

The ARIMA and regression with ARIMA errors were engaged to investigate Somaliland's monthly CPI from January 2013 to April 2020, after applying the Box-Jenkins methodology and regression analysis. The study mainly aimed to forecast the monthly CPI in Somaliland for the May 2020 – April 2021 period and selected the best fitting model based on

how well the model captures the stochastic variance in the data. It was obtained that the ARIMA (0, 1, 3) model is a reasonable and acceptable model for forecasting Somaliland's CPI in the next twelve months. Generally, CPI in Somaliland has shown an upward trend over the forecasted period. Based on the findings of the study, policymakers in Somaliland should pursue more sensible monetary policies to combat such an increase in inflation as reflected in the forecasts. In this respect, fiscal and monetary authorities are advised to take tight economic policy measures to tackle the inflation threat in Somaliland.

References

- [1] Manga, G. S. (1977). *Mathematics and Statistics for Economics*, Vikas Publishing House, New Delhi.
- [2] Subhani, M. I., and Panjwani, K. (2009). Relationship between Consumer Price Index (CPI) and Government Bonds, *South Asian Journal of Management Sciences*, Vol. 3, No. 1, pp. 11–17.
- [3] Costa, D. L. (2001) Estimating Real Income in the United States from 1888 to 1994: Correcting CPI Bias Using Engel Curves. *Journal of Political Economy*, Vol. 109, pp. 1288-1310.
- [4] Central Statistics Department. (2020). *Consumer Price Index Publication*. Ministry of National Planning and Development, Hargeisa, Somaliland.
- [5] Isakova, A. (2007). Modeling and Forecasting inflation in developing Countries: The case of Economies in Central Asia. Center for Economic Research and Graduate Education, Discussion Paper, (2007-174).
- [6] Norbert, H., Wanjoya, A., and Waititu, A. (2016). Modeling and forecasting consumer price index (Case of Rwanda). *American Journal of Theoretical and Applied Statistics*, Vol. 5, pp. 101-107.
- [7] Montgomery, D. C., Jennings, C. L., and Kulahci, M. (2008). *Introduction to Time Series Analysis and Forecasting*. New York: John Wiley & Sons.
- [8] Brocwell, P. J., and Davis, R. A. (2002). *Introduction to time series and Forecasting*. New York: Springer.
- [9] Xiao, Z. (2016). CPI Prediction Based on ARIMA Model. In *7th International Conference on Education, Management, Information and Computer Science (ICEMC 2017)*. Atlantis Press.
- [10] Chatfield, C. (2004). *The Analysis of Time Series an Introduction Sixth Edition*. New York: Chapman & Hall/CRC.
- [11] Wei, W. S. (2006). *Time Series Analysis Univariate and Multivariate Methods Second Edition*. Canada: Pearson education Inc.
- [12] Cryer, J. D., and Chan K. S. (2008). *Time Series Analysis with Application in R*. New York: Springer.
- [13] Box, G. E. P., and Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. Holden-Day, San Francisco.
- [14] Adam, S. O., Awujola, A., and Alumgudu, A. I., (2014). Modeling Nigeria's CPI using ARIMA model. *Int. Journal of Development and Economic Sustainability*, Vol. 2, No. 2, pp. 37-47.
- [15] Hamid, S. A., and Dhakar, T. S. (2008). The behaviour of the US consumer price index 1913–2003: a study of seasonality in the monthly US CPI. *Applied Economics*, Vol. 40, No. 13, pp. 1637-1650.
- [16] Zhang, F. W., Che, W. G., Xu, B. B., and Xu, J. Z. (2013). The Research of ARMA Model in CPI Time Series. In *Applied Mechanics and Materials* (Vol. 347, pp. 3099-3103). Trans Tech Publications Ltd.
- [17] Kharimah, F., Usman, M., Widiarti, W., and Elfaki, F. A. M. (2015). Time series modeling and forecasting of the consumer price index Bandar Lampung. *Science International Lahore*, Vol. 27, No. 5, pp. 4619-4624.
- [18] Gujarati, D. (1995). *Basic Econometrics*. McGraw-Hill Inc., Singapore.
- [19] Ljung, G. M., and Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, Vol. 65, No. 2, pp. 297-303.
- [20] Nyoni, T. (2018). Box –Jenkins ARIMA Approach to Predicting net FDI inflows in Zimbabwe, Munich University Library – Munich Personal RePEc Archive (MPRA), Paper No. 87737.