IMPACT: International Journal of Research in Business Management (IMPACT: IJRBM)
ISSN (P): 2347-4572; ISSN (E): 2321-886X
Vol. 5, Issue 4, Apr 2017, 45-52

APPLICATION OF ARIMA MODEL FOR PREDICTING CASHEW NUT PRODUCTION IN INDIA – AN ANALYSIS

E. ELAKKIYA[1], M. RADHA[2] & R. SATHY[3]

[1]Research Scholar, Department of Physical Sciences and Information Technology, Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu, India

[2]Assistant Professor, Department of Physical Sciences and Information Technology, Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu, India

[3]Professor, Department of Physical Sciences and Information Technology, Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu, India

ABSTRACT

A statistical modeling approach (Box-Jenkins' ARIMA model) has been used for the study to predict Cashew nut production in India. The order of the best ARIMA model was found to be (2,1,1). Further, efforts were made to predict, as accurate as possible, the future cashew nut production for a period up to five years of fitting ARIMA (2,1,1) model to our time series data. The prediction results were shown that the annual cashew nut production to grow in 2016, then its take a sharp dips in 2015 and in subsequent years 2016 through 2020.

KEYWORDS: ARIMA Model, Cashew Nut, Data and Predicting

1. INTRODUCTION

Cashew (*Anacardium occidentale*) was originated in Brazil. In the 16th century, Cashew was introduced for aforestation and soil conservation in India. It was highly commercialized in eight states such as Andhra Pradesh, Goa, Karnataka, Kerala, Maharashtra, Orissa and Tamil Nadu. The area of growing cashew was 7.30 lakh ha and annual production of about 4.60 lakh tonnes of raw cashew nut.

In this paper, an effort is made to predict the production of cashew for five successive years. Box and Jerkin (1960) developed an Autoregressive Integrated Moving Average (ARIMA) model for predicting the production. This model is used to forecast a single variable. The important reason for selecting an ARIMA model in this study is taken into account the non-zero autocorrelation between the successive values of the time series data.

The formation of the ARIMA model depends on the characteristics of the series. Table 1 represents the 25 years' cashew nut production in India. The data are taken from the secondary source from the Directorate of Cashew and Cocoa Development, in India from 1990 to 2015. In this study, to use for GRETL (Gnu Regression, Econometrics and Time-series Library) software, SPSS and EXCEL for plotting the graphs and analysis of the dataset.

2. MATERIALS AND METHODS

The formulation of the ARIMA model depends on the characteristics of the time series. The Autoregressive Integrated Moving Average (ARIMA) model is the forecasting model which is popularized by Box and Jerkins (1976). An ARMA (p, q) model is a combination of Autoregressive (AR) which indicates that there is an association among present and past values, a random value and a Moving Average (MA) model is a linear combination of

white noise error terms. The ARMA (p, q) process can be defined as follows;

Autoregressive Model

The notation AR (p) refers to the autoregressive model of order p. The AR (p) model is written as

$$Yt = C + \sum_{i=1}^{p} \varphi i Yt-i + \varepsilon_t, \quad \textbf{(1)}$$

Where $\varphi 1$, …,are parameters, C is a constant and the random variables $\varepsilon t$is white noise.

2.2. Moving – Average Model

The notation MA (q) refers to the moving average model of order q.

$$Yt = \mu + \varepsilon t + \sum_{i=1}^{q} \theta i \varepsilon_{t-i}, \; t = 1,2 \ldots \quad \textbf{(2)}$$

Where $\theta 1,\ldots,\theta_q$are parameters, μ is the expectation of $Y_t$and $\varepsilon t$is white noise (error term).

2.3. ARMA (p, q) Model

The notation ARMA (p, q) refers to the model with p autoregressive terms and q moving- average terms. Finally, by combining equation 3.01 and 3.02 the ARMA (p, q) is given by

$$Yt = C + \varepsilon t + \sum_{i=1}^{p} \varphi i + \sum_{i=1}^{q} \theta i \varepsilon t-i \quad \textbf{(3)}$$

To achieve the above mentioned, there are three steps namely;

- Model identification
- Model estimation
- Model verification or diagnostic checking.

2.4. Model Identification

Model identification was used for evaluate the different models for given data. This step helps to determine the value of p, d and q. Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) was used for determine the value of p, d and q which can be done by observing the graph of the data or autocorrelation, partial autocorrelation functions (Makridakis et al., 1998). For any ARIMA (p, d, q) process, the theoretical (PACF) has non-zero partial autocorrelation at 1, 2… p and has zero partial autocorrelations at all lags. The non-zero lags of the sample PACF & ACF are tentatively accepted as the p and q parameters. For a non-stationary series, the log data is differenced to make the series stationary. The number of times the series differenced determines the order of d. Thus, for a stationary data d=0 &ARIMA (p, d, q) can be written as ARMA (p, q).

2.5. Model Estimation

An optimal model has been identified; the model estimation methods make it possible to estimate simultaneously all the parameters of the process, the order of integration coefficient and parameters of an ARMA structure. The estimates of the exact maximum likelihood proposed by so well are the vector $\hat{\beta} = (\hat{d}, \hat{\varphi}, \hat{\theta})$ which maximizes the log -

likelihood function L (β).

$$\text{“L }(\beta) = -(\frac{n}{2}) \text{ In } (2\pi) - (\frac{1}{2}) \text{ In (R)} - (\frac{1}{2}) X'R^{-1}x\text{”} \quad \textbf{(4)}$$

Where R is the variance – covariance matrix of the process

2.6. Model Verification

The model verification is the last step which helps to identify the residuals of the model to decide whether to accept or reject the model. For methods of residual assessment, if evidence in letter case, due to inadequacies of the model. Therefore, there is need of the repetition of step 2 or even step 1. Thus the model building is an iterative, interactive process. So, given multiple competing models, we decide upon a final one model which is one popular method to use a model selection criterion; Akaike's Information Criterion (AIC) Schwartz information Criterion (SIC) and Hannan Quinn Criterion (HQC) which attempts to choose a model that adequately describes the data but in the most parsimonious way possible, or in other words, minimizing the number of parameters.

2.6.1. Akaike's Information Criterion

Akaike's Information Criterion (AIC) originally proposed by Akaike, attempts to select a good approximating model for inference based on the principle of parsimony. AIC proposes the use of the relative entropy, or the Kull back – Leibler (K-L) information as a fundamental basis for model selection. A suitable estimator of the relative K-L information is used and involves two terms. The first term is a measure of lack of model fit, while the second is a “penalty” for increasing the size of the model, assuring parsimony in the number of parameters. The AIC criterion to be minimized is

$$\text{AIC (n)} = \log (\delta_q^2) + \frac{2n}{T} \quad \textbf{(5)}$$

Where n is the dimensionality of the model, δ_q^2 is the maximum likelihood estimate of the white noise variance, and T is the sample size.

2.6.2. Schwarz's Bayesian Information Criterion

The Bayesian Information Criterion (BIC), originally proposed by Schwarz was derived in a Bayesian context and is “dimension consistent” in that it attempts to consistently estimate the dimension of the true model. It assumes a true model exists in the set of candidate models, therefore requires a large sample size to be effective. The BIC Criterion to be minimized is

$$\text{BIC (n)} = \log (\delta_q^2) + \frac{n \log (T)}{T} \quad \textbf{(6)}$$

Where n is the dimensionality of the model, δ_q^2Is the maximum likelihood of the estimate of the white noise variance, and T is the sample size.

2.6.3. Hannan – Quinn Criterion

The Hannan – Quinn (HQ) Criterion, originally proposed by Hannan & Quinn was derived from the law of the iterated logarithm, it is another dimension consistent model and only differs from AIC and BIC with respect to the “penalty term”. The HQ Criterion to be minimized is

$$HQ(n) = \log(\delta_q^2) + \frac{2n\log(T)}{T} \quad (7)$$

Where n is the dimensionality of the model, δ_q^2Is the maximum likelihood of the estimate of the white noise variance, and T is the sample size. Hannan and Rissanen later replaced the term log (n) with log in to speed up the convergence of HQ.

2.7. Tests for Stationarity

First, we have to test the stationary of the time series. We can use the formal and the most popular method to test the stationary of a series is the unit root test. This test is used for identify the order of integration of non-stationary variable, so there may be difference before being included in the regression equation. The Augmented Dickey Fuller (ADF) test is the most commonly used unit root test.

2.7.1. Augmented Dickey Fuller Test

The test was first introduced by Dickey and Fuller (1979) to test for the presence of unit root(s). The regression model for the test is given as

$$\Delta y_t = \gamma y_{t-1} + \beta x_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \Delta y_{t-2} + \ldots + \delta_p \Delta y_{t-p} + \varepsilon_t \quad (8)$$

The hypothesis testing is $H_0 : \gamma = 0$ (the series contain unit root(s)) $H_1 : \gamma < 0$ (the series is stationary)

Test statistic, $t_\gamma = \frac{\gamma}{SE(\gamma)}$, (9)

Δy_t = the difference series y_{t-1} = the immediate previous observation δ_1,…., = the coefficient of the lagged difference term up to p x_t = the optimal exogenous regresses which may be constant the constant or constant trend= parameters to be estimated.

3. RESULTS AND DISCUSSIONS

3.1. Time Series Analysis and Building ARIMA

The given set of data in *Table 1* is used to develop a forecasting model. The *Figure 1represents the line plots of cashew nut production in India.*

According to materials and methods, the ARIMA model included the following four steps for forecasting. There are (1) Model Identification, (2) Parameter Estimation and Selection, and (3) Diagnostic Checking we can (4) Model for forecasting application

3.2. Model Identification

The first stage of model building is to detect whether the variable is stationary or not. The time plot of the cashew nut production data in *Figure 1clearly shows that the data is not stationary*. We obtained a time series of first order differencing and Figure 2 below is the line plot of the first order differenced cashew nut production data.

3.3. Test for Stationarity: Augmented Dickey-Fuller (ADF) Test

First order differencing (d=1) means we generate a table of differenced data. The ADF test result, as obtained upon application, is shown below:

Dickey-Fuller = -3.73283, Lag order = 3, p-value = 0.0037

Therefore, fail to accept the H_0 and hence can conclude that H_1 is true, i.e. the series is stationary in its mean and variance. Thus, there is no need for further differencing the time series and we adopt d = 1 for our ARIMA (p, d, q) model.

3.4. Correlogram and Partial Correlogram

Figure 3 represents that the plot of correlogram (ACF) for lags 1 to 20 off the first order difference time series and partial correlogram (PACF) for lags 1 to 20 of the differenced time series of the cashew nut production in India.

The correlogram infers (Figure 3) that the lag 1 does exceed the significance limits and auto-correlations tail off to zero after lag 3. Although the autocorrelation at lag 5 almost touching the significant limits, rest all coefficients between lag 6 and 20 are well within the limits. The partial correlogram, Figure 3, infers that the partial auto-correlation coefficient does exceed significant limits at lag 1 and after lag 2 partial autocorrelation tail off to zero. All the other PACFs from lag 2 to 20 are within the significant limits. Table 2 represents that the ACF and PACF coefficients for lag 1 to 20 of that first order differenced series.

Define the following possible ARMA (autoregressive moving average) models for the first -differenced time series data of cashew nut production in India:

- An ARMA (2,0) model, i.e. AR model of order p=2 since the partial autocorrelation is zero after lag 2 and the autocorrelation is zero.
- An ARMA (0,3) model, i.e. MA model of order q=3 since the partial autocorrelation is zero and the autocorrelation is zero after lag 3.
- An ARMA (p,q) model i.e. A mixture model with p and q both greater than 0 since autocorrelation and partial autocorrelation both tail off to zero.

3.5. Selecting the Candidate Model for Forecasting

Since ARMA (2,0) has 2 parameters in it, ARMA (0,3) has 3 parameters in it and ARMA (p, q) has at least 2 parameters in it. In the next step, we have to device the best ARIMA model using the ARMA (2, 1) model (with p=2 & q=1). Therefore, based upon the conditions, we can have only followed three tentative ARIMA (p, d, q) models:

ARIMA(p,d,q): ARIMA(2,1,0), ARIMA(2,1,1), and ARIMA(2,1,2)

To select as the best suitable model for forecasting out of three above, we will choose the one with the lowest BIC & AIC values. Following *Table 3 summarizes the output of each of the fitted ARIMA models in our time series* (of cashew nut production data):

We can clearly observe in the table above that the lowest AIC and BIC values are for the ARIMA (2,1,1) model with (p=2, d=1 and q=1) and hence this model can be the best predictive model for making forecasts of future values of our time series data.

3.6. Forecasting Using Selected ARIMA Model

To select the model (*Table 3*) ARIMA (2,1,1), which we are fitting to time series data for fitting ARMA (2,1) model of first-order difference to our time series. Also, ARMA (2,1) model, which has two parameters in it, can be rewritten as an AR model of order 2 and an MA model of order 1. Now to fit the chosen ARIMA (2,1,1) model to forecast for the future values of our time series. Table 4 clearly shows that the forecast for the next 5 years with 80%, 95% and 99% (low and high) prediction internals:

Figure 4 shows the plot for 5- year forecast of the cashew nut production by fitting ARIMA (2,1,1) model to our time series data:

To investigate the distribution of forecasting errors, we will plot the errors (standard residuals). *Figure 5 (a), 5 (b) and 5 (c) below show various plots and histograms of standard residuals (forecast errors) of fitted ARIMA (2,1,1) model*:

To investigate further whether there are any correlations between successive forecast errors, to plot the ACF& PACF of the forecast errors. Following Figure 6 represents ACF and PACF of the forecast errors:

All the ACFs & PACFs of residuals of fitted ARIMA for lag 1 to lag 20 are within the significant limits. This means ACF and PACF concluded that there are no non-zero autocorrelations in the forecast residuals (or standard errors) at lag 1 to 20 in the fitted ARIMA (2,1, 1) model. The Box-Ljung test results are shown in the *table 5 shows, the Box-Ljung test statistics* while below represent the plot of the Box-Ljung p-values for the fitted model:

The statistics and large p-values in both the tests above are suggesting us to accept the null hypothesis that all of the autocorrelation functions in lag 1 to 20 are zero.

4. CONCLUSIONS

From the study, the ARIMA (2, 1, 1) was the best candidate model selected for making predictions for up to 5 years for the cashew nut production in India using a 25 years' time series data. ARIMA model was used for the reasons of its capabilities to make predictions using a time series data with any kind of pattern and with autocorrelations between the successive values in the time series. The study also statistically tested and validated that the successive residuals (forecast errors) in the fitted ARIMA model were not correlated and the residuals seem to be normally distributed with mean zero and constant variance. Hence, conclude that the selected ARIMA (2, 1,1) seem to provide an adequate predictive model for the cashew nut production in India.

REFERENCES

1. Contreras, J, Espinola, R, Nogales,F.J, Conejo, A.J. 2003.ARIMA Models to Predict Next Day Electricity Prices.IEEE transactions on power system.Vol (1)8, No.3, pp. 1014-1020.
2. G.C.1995. Yield Gap and Constraints in Technology Adoption of Cashew Nut Cultivation in the Konkani Region of Maharashtra, The cashew.9 (4); 13-7.
3. Haldankar; P.M. Chavan (2004), Strategies and Constraints for Cashew Production in Maharashtra. The Cashew 18 (2); 13-115.
4. JenkinsMethodology".Applied Econometrics(Second Ed.). Palgrave.

5. SK, Nag, et al. "Allotment Procedure of Cashew Nut Trees in Bastar District of Chhattisgarh State: A Case Study." (2016).
6. Jiban Chandra Paul. ShahidulHoque. Mohammad MorshedurRahman. 2013. Selection of Best ARIMA Model for Forecasting Average Daily Share Price Index of Pharmaceutical Companies in Bangladesh, Glo.J. Inc.Vol(13); 3. ISSN:2249- 4588.