

**FORECASTING OF RICE PRODUCTION IN INDIA: AN
APPLICATION OF BOX-JENKINS APPROACH**

*A Dissertation Submitted to the Department of Statistics, Gargaon College for the Degree
of Bachelor of Science in Statistics Awarded by Dibrugarh University*



Submitted by

Lisbon Borgohain

B.Sc. Sixth Semester, 2023

Roll No: 13220139 Registration No: S2006883

Supervised by

Dr. Bornali Dutta

Assistant Professor,

Department of Statistics

DEPARTMENT OF STATISTICS

GARGAON COLLEGE, SIMALUGURI - 785686, ASSAM


GARGAON COLLEGE

P.O. : SIMALUGURI; DIST.: SIVASAGAR; ASSAM - 785686
Email ID: statistics.gargaoncollege2020@gmail.com

CERTIFICATE

This is to certify that, Lisbon Borgohain, a student of B. Sc. Sixth Semester, having Hons in Statistics, has prepared this Project Report entitled, "*Forecasting of Rice Production in India: An Application of Box-Jenkins Approach*", under my guidance and supervision. It embodies the result of her own investigation.

Date: 12/06/2023
Place: Gargaon college


12/6/23
[Dr. Bornali Dutta]
Assistant Professor
Department of Statistics,
Gargaon College

CHAPTER-I

INTRODUCTION

1.1 Aim of the field work:

Any scientific study is not perfect and acceptable without any practical work with suitable result or experimental verification. One must have practical experience in addition to the theoretical knowledge. Project or field work is a part of practical experience. Field research or fieldwork is the collection of information outside a laboratory or workplace setting. The approaches and methods used in field research vary across disciplines. For, biologists who conduct field research may interview or observe people in their natural environments to learn their languages, social structures or to study any socio-economic situations. Field research involves a range of well-defined, although varied methods: informal interviews, direct observation, participation in the life of the group, collective discussions and analyses of personal documents produced within the group, collection discussion, and analyses of personal documents produced within the group, self-analysis, results from activities undertaken off or on-line and life-histories. Although the method generally is characterized as qualitative research it includes quantitative dimensions also.

The project work makes one's mind research oriented and gives confidence to take any independent study in future where a researcher can take an appropriate decision on the basis of the findings he acquires through his observation and analysis. Any research project or field work study is powerful and acceptable if the result can be expressed or communicated in numbers, i.e., based statistically. The numbers and testing of the numbers to find real "significant" differences in populations is a necessary form of communication for scientists. Use of numbers also provides an air of credibility to scientific studies. Use of statistics requires planning and statements about experimental design, methods and data base structure. The numbers also force one to formalize his or her thinking about hypotheses and how best to search out meaning from the data. When it comes to scientific research, the goal of one's work is through the use of data to converge on principles, laws and data relationships that inform the researcher about how nature operates. Numbers and statistics help to measure this progress and to heighten the confidence in the results.

1.2 Collection of data:

It is very essential to collect data for applying statistical methods to as type of enquiry, whether it is in business /commerce or economics or social science. The numerical facts or measurements obtained in the course of enquiry into phenomenon marked by uncertainty constitute statistical data. The basic problem is to collect facts and figures relating to the particular phenomenon under study. In case, the collected data are inaccurate and unreliable, the further analysis of data would result in wrong conclusion. Therefore, proper method should be employed in the collection of data for studying the true behaviour of variables under study. The collection of data is the foundation upon which the whole superstructure of statistical analysis is to be raised. One should know very clearly the statistical unit e. g. population, sampling units, measurements to be taken, sample size etc before collecting the data.

A fundamental question to be considered at the outside is whether the collection of data should be done by complete enumeration or by sampling. In the former case, each and every of the group to which the data are to relate is covered and information gathered for everyone separately. In the latter, only some individual forming a representative part of the group are covered, either because the group is too large or the items on which information is sought are too numerous. Complete enumeration may lead to greater accuracy and greater refinement in analysis, but it may be very expensive and time-consuming operation. A sample designed and taken with care can produce results may be very sufficiently accurate for the purpose of the enquiry and it can save time and money.

The information sought may be gathered from the individuals by one of the three methods: the questionnaire method, the interviewer method and the method of direct observation. In economic and social enquires, information is almost always collected by having someone to fill up a form or questionnaire. It should be decided whether the forms should be completed by an enumerator or investigator who collects data by asking questions and noting down answers, or whether this should be left with the respondent to be filled up on his own. In the interviewer method, enumerator goes from one informant to another and elicits the required information. This method is used to population census. In the method of direct observation, the enquire gets the data directly from the field of enquiry without having to depend on the informants.

Great care is to be taken in drafting a questionnaire or schedule, as this is the medium through which information is collected. Further it is also to be seen that the information collected is usable. Apart from care, expertise such as skill, wisdom, experience of the phenomenon under enquiry is needed in drafting a questionnaire or schedule. Though there are no hard and fast rules for designing a questionnaire, there are a few general points to be kept in mind-

- (i). The questions put should be clear, concise and unambiguous.
- (ii). Delicate questions are to be put with great care- often indirect questions should be put to get answer to some pertinent point. It is sometimes desirable to avoid very delicate questions.
- (iii). The size of the questionnaire or schedule should be small. It saves time, both for the enumerator and the respondent.
- (iv). There should be a natural, logical order in which questions are put.
- (v) It should be noted that the information collected through questions should be such that it is usable.

1.3 Present Study:

The present study is very modest one undertaken with limited scope, time and resources coupled with inexperience. **The Project Work is a part of the curriculum in the B. Sc. Statistics (Hons) Examination of the Sixth Semester CBCS Under-Graduate Course under Dibrugarh University.**

The title of the Project Dissertation is: "**Forecasting of Rice Production in India: An Application of Box-Jenkins Approach**". The data of the present study are secondary and obtained from United States Department of Agriculture consisting of yearly production of rice in 1,000 metric tons for the period of 1960-2022, a total of 63 data points.

In this study, Box-Jenkins methodology is used to predict the yearly production of rice. Model adequacy is examined using AIC(Akaike Information Criterion), AICc(Corrected Akaike Information Criterion), BIC(Bayesian Information Criterion), RMSE(Root Mean Square Error), MASE(Mean Absolute Scale Error), MAPE (Mean Absolute Percentage Error) etc. It is a combination of five chapters. The chapter I, i.e., the present chapter is an introductory one, where the aim of the field work and data collection methods are discussed followed by rice production data in India. Chapter-III covers a discussion of importance of

time series analysis, various components of time series, details of Box-Jenkins Methodology and different accuracy measures are discussed. Chapter-IV comprises the results and interpretation of the study and final comments are made in Chapter-V. It concludes with a list of references without which this work could not be completed.

CHAPTER – II

Forecasting of Rice Production in India: An Application of Box-5Jenkins Approach

Agriculture is the backbone of Indian's economy. It employs nearly half of the workforce in the country. Besides providing for the livelihoods, it also addresses food security of the country. It contributes to 17.5% of G.D.P. Population of India increasing at an alarming rate. The country's requirement for food grains in order to provide for its population is projected to be 300 million tonnes by 2025. To feed these people, the planners should have an idea about the forecasting of production of major crops. To improve the overall food security situation in India, forecasting of any crop is an important area of research for management of agriculture, formulation and implementation of policies related to prices, export-import decision and development of modern technology to improve the production.

Rice is one of the most prominent crops of India as it is the staple food of nearly 70% of India's population. It is the backbone of livelihood for millions of rural people and plays a vital role in country's food security. Rice provides 27 % of dietary energy supply, 20% of dietary protein and 3 % of dietary fat. Rice can contribute nutritionally significant amounts of thiamine, riboflavin, niacin and zinc to the diet and also smaller amounts of other micronutrients. India occupies an important position both in area and production of rice. To meet the population demand, it is necessary to increase the productivity per unit of area of rice with enhanced resource use efficiency. Now-a-days, arable land degraded day by day. Major constraint for productivity and sustainability of rice -based system in the country are the inefficient use of inputs like fertilizer, water, labour, increasing scarcity of water, emerging challenges from climate change, rising fuel prices, urbanization, environmental pollution, less liking for agricultural work by youth.

Rice is a cereal grain and monocot; a plant with a seed that has one embryonic leaf. The only two types of cultivated rice are African rice (*Oryza glaberrima*) and Asian rice (*Oryza sativa*). The plant itself grows between 90- 150 cm long and 15 mm wide. The small flowers have 6 anthers (the part of the stamen with pollen) and 2 stigmas (where pollen germinates). It has a dry fruit and spreads its seed through the wind. The grain gets processed into rice. It's spread across warm, tropical and aquatic conditions like flood plains, wetlands, ponds and

streams. Whilst rice farms are global, it's concentrated mainly in Asian developing countries. But it needs a good infrastructure to support the industry, including disease and pest control. Rice can take up to 200 days to mature and then it's a hard process of manual work to hand-harvest it from the paddy fields and dries out the plants. Then the seeds are threshed and milled with a huller, removing the outer husk until it becomes rice. However, it loses some of its nutritious properties in the process. Perennial wild rice still grows in Assam and Nepal. It seems to have appeared around 1400 BC in southern India after its domestication in the northern plains. It then spread to all the fertile alluvial plains watered by rivers. Climatic requirements in India rice are grown under widely varying conditions of altitude and climate. Rice cultivation in India extends from 8 to 35°N latitude and from sea level to as high as 3000 meters. Rice crop needs a hot and humid climate. It is best suited to regions which have high humidity, prolonged sunshine and an assured supply of water. The average temperature required throughout the life period of the crop ranges from 21 to 37° C. Maximum temperature which the crop can tolerate 400 C to 420 C. Rice is a nutritional staple food which provides instant energy as its most important component is carbohydrate (starch). On the other hand, rice is poor in nitrogenous substances with average composition of these substances being only 8 per cent and fat content or lipids only negligible, i.e., 1 percent and due to these reasons, it is considered as a complete food for eating. Rice flour is rich in starch and is used for making various food materials. It is also used in some instances by brewers to make alcoholic malt. Likewise, rice straw mixed with other materials is used to produce porcelain, glass and pottery. Rice is also used in manufacturing of paper pulp and livestock bedding. The variability of composition and characteristics of rice is really broad and depends on variety and environmental conditions under which the crop is grown. In husked rice, protein content ranges in between 7per cent to 12per cent. The use of nitrogen fertilizers increases the percentage content of some amino acids. Ancient Ayurvedic literature testifies the medicinal and curative properties of different types of rice grown in India. Rice is mainly grown in two types of soils i.e., (i) uplands and (ii) lowlands. The method of cultivation of rice in a particular region depends largely on factors such as situation of land, type of soils, irrigation facilities and availability of labour intensity and distribution of rainfalls. Rice provides 21% of global human per capita energy and 15% of per capita protein. Although rice protein ranks high in nutritional quality among cereals, protein, content is modest. Rice also provides minerals, vitamins, and fibre, although all constituents except carbohydrates are reduced by milling. Moreover, India has the largest area under rice cultivation as it is one of the principal food crops. It is, in fact, the dominant crop of the country.

Rice can be cultivated by different methods based on the types of regions. But in India, traditional methods are still in use for harvesting rice. The fields are initially ploughed, and fertilizer is applied which typically consists of cow dung, and then the field is smoothed. The seeds are transplanted by hand and then through proper irrigation, the seeds are cultivated. Rice grows on a variety of soils like silts, loams and gravels. It can tolerate alkaline as well as acid soils. However clayey loam is well suited to the raising of this crop. Actually, the clayey soil can be easily converted into mud in which rice seedlings can be transplanted easily. Proper care has to taken as this crop thrives if the soil remains wet and is under water during its growing years. Rice fields should be level and should have low mud walls for retaining water. In the plain areas, excess rainwater is allowed to inundate the rice fields and flow slowly. Rice raised in the well-watered low land areas is known as lowland or wet rice. In the hilly areas, slopes are cut into terraces for the cultivation of rice. Thus, the rice grown in the hilly areas are known as dry or upland rice. The yield of upland rice per hectares is comparatively less that of the wet rice.

Among the rice growing countries in the world India has the largest area under rice crop and ranks second in production next to China. It is important cereal crop of India and is cultivated on about 45.54 m area with an annual production of 99.18 million tones and a productivity of 1460 kg/ha. Among the rice producing states of India Tamil Nadu ranks sixth in Production (5.67 million Tonnes) and second productivity of 3070 kg/ha with an area of 1.85m ha.

CHAPTER-III METHODOLOGY

3.1 A brief introduction about Time Series:

A time series is a set of observations measured at time or space intervals arranged in chronological order i.e., in accordance with its time of occurrence. It reflects the dynamic pace of movements of a phenomenon over a period. Most of the series relating to Economics, Business and Commerce, e.g. the series relating to prices, production and consumption of various commodities; agricultural and industrial production, national income and foreign exchange reserves, investment, sales and profits of business houses; bank deposits and bank clearings, prices and dividends of shares in a stock exchange market etc are all time series spread over a period of time.

Mathematically, a time series is defined by the functional relationship

$Y=f(t)$ where y is the value of the phenomenon (or variable) under consideration at time t . For example

- (i). The sale (y) of a departmental store in different months (t) of the year.
- (ii). The temperature (y) of a place on different days (t) of the week and so on constitute time series.

3.1.1. Importance of time series analysis:

The analysis of time series is useful to economists and business persons and to scientists, sociologist etc. It has also found its utility in meteorology, seismology, oceanography, geomorphology etc in earth sciences, electrocardiograms, electroencephalograms in medical sciences and problem of estimating missile trajectories. Time series analysis helps in understanding the following phenomena.

- (i). It helps in knowing the real behaviour of the past.
- (ii). It helps in predicting the future behaviour like demand, production, weather conditions, prices etc.
- (iii). It helps in planning the future operations.

(iv). Analysis of time series helps to compare the present accomplishments with the past performances.

(v). Two or more time series can be compared belonging to the same reference period.

3.1.2 Drawbacks of the Time Series Analysis:

(i). The conclusions drawn based on time series analysis are not cent percent true.

(ii). Time series analysis is unable to fully adjust the influences affecting a time series like customs, climate, policy changes etc.

(iii). The complex forces affecting a time series existing at certain period may not be having the same complex forces in future. Hence, forecast may not hold true.

3.2 Components of Time Series:

Through the detailed study of the time series, we can extract an idea about the changes in the effects of various factors which interact simultaneously. These effects are classified in some major categories. These components are known as the components of time series. The components are

Secular trend:

The word trend means the tendency. So, secular trend is that component of the time series which gives the general tendency of the data for a long period. It is smooth, regular and long-term movement of a series. The steady growth of the sale status of a particular commodity of a company or the fall of demand for a certain article for long years can be studied through this secular trend.

Seasonal variation:

Short-term fluctuations observed in a time series data, particularly in a specified period usually within a year are called seasonal variations. For instance, certain items have more sales in a particular season like ice cream in summer, rain coats in rainy season and woollens in winter season. Similarly, first week of a month records greater sale of grocery than the last week of a month. All such variation in a time series comes under seasonal variation. Seasonal variations are more akin to climatic and weather conditions.

Cyclical Variation:

The oscillatory movements in a time series with period of oscillation greater than one year are termed as cyclical variations. These variations in a time series are due to ups and downs recurring after a period greater than one year. The cyclical fluctuations, though regular, are not necessarily uniformly periodic, i.e. they may or may not follow exactly similar patterns after equal intervals of time. A complete cycle usually has four constituents namely, prosperity, recession, depression and recovery.

The study of cyclical variations is of great importance to business executives in the formulation of policies aimed at stabilizing the level of business activity. Knowledge of the cyclic component enables a businessman to have an idea about the periodicity of the booms and depressions and accordingly he can take timely steps for maintaining stable market for his product.

Irregular Variation:

Irregular variations are those which are quite unpredictable. The effects due to flood, draughts, famine, devastating storms, earthquake or any natural calamities are known as irregular variations. The variations which include other than trend, seasonal and cyclical variations are all irregular.

3.3 Mathematical models for time series:

The following are the two models commonly used for the decomposition of a time series into its components.

3.3.1. Additive Model or Decomposition by additive Hypothesis:

According to the additive model, the time series can be expressed as

$$Y_t = T_t + S_t + C_t + I_t$$

Where Y_t is the time series value at time t and T_t, S_t, C_t, I_t represents the trend, seasonal, cyclical and random variations at time t .

The additive model assumes that all the four components of the time series operate independently of each other so that none of these components has any effect on the remaining three. This implies that the trend, however, fast or slow it may be, has no effect on the

seasonal and cyclical components; nor do seasonal swings have any impact on cyclical variations and conversely.

3.3.2. Multiplicative Model or Decomposition by Multiplicative Hypothesis:

In a traditional or classical time series analysis, the most assumed mathematical model is the multiplicative model. Here it is assumed that any observation Y at time t is because of the product of the effect of the four components of a time series namely, Trend(T), Seasonal Variation (S), Cyclical Variation (C) and the Irregular Variation (I), i.e.,

$$Y = T \times S \times C \times I$$

Further the multiplicative model does not assume the independence of the four components of the time series. It is appropriate for projections.

3.3.3 Mixed Model:

A mixed model is a mathematical relation which is expressed as a combination of multiplicative and additive components of a time series. They can be combined in several ways. Such types of models are hardly used. Some of the examples of mixed models are given below:

$$Y = T + S \times C + I$$

$$Y = T + S \times C \times I$$

$$Y = T + S + C + I$$

3.4 Testing of Stationarity of a Time Series:

When there is a trend or seasonality present in time series data, then it is not a stationary one but to do something with the help of time series, it must be stationary one. Now the problem is to find out whether the time series is stationary or not. There are various methods for testing stationarity of a time series

(i) Q –Statistic:

Box and Pierce developed the Q-statistic to check the stationarity of a time series. The test statistics is defined as

$$Q = n \sum_{k=1}^m \hat{\rho}_k^2 \quad (3.4.1)$$

Where, n = number of observations used to fit the model, m = lag length, ρ_k = autocorrelation coefficient.

The Q statistic is approximately distributed as χ^2 distribution with m degrees of freedom. The test may reject the null hypothesis if the calculated value exceeds the critical value at desired level of significance and it may be concluded that the series is stationary.

(ii) Correlogram:

Correlogram is an important tool to check the nature of the time series. To check the stationary of the series one should plot the autocorrelations (ρ_k) against their lag (k). If the values of the autocorrelations are dense or cluster around the straight line, then the time series is stationary. On the other hand if the values of the autocorrelation move within the limit -1 and +1 and not cluster around the straight line then the time series is considered as non stationary.

(iii) Augmented Dickey-Fuller Test:

One of the disadvantages of Dickey-Fuller test is that the error term is non auto correlated. But, in practical purpose the errors are considered as correlated. Moreover, the ADF test assumes that the series is non-stationary. The alternative hypothesis differs depending on which version of the test is used, but it is usually stationary. It is an augmented version of the Dicky-Fuller test for a larger and more complicated set of time series models.

3.5 Model Selection Criteria in Time Series Analysis:

After estimating the parameters of the proposed models the next step is to choose the best model using some suitable model selection criterion. For example: Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), Corrected Akaike Information Criterion (AIC_C) which are measures of goodness of fit and they are generally used to select the best model. Details of these measures are discussed below:

(i) Akaike Information Criterion:

It was formulated by the Japanese Statistician Hirotugu Akaike in the year 1974. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Mathematically, AIC is defined as follows-

$$AIC = \ln \left(\frac{\sum_{i=1}^n \epsilon_i^2}{n} \right) + \frac{2\lambda}{n} \quad (3.5.1)$$

Where ε_t represents the error term, n = number of observations and λ = number of parameters

(ii) Bayesian Information Criterion:

Bayesian Information Criterion (BIC) is also a broadly used model selection principle among a finite set of models. It was developed by Schwarz (1978), Raftery (1995). Mathematically, BIC is calculated by the equation-

$$\text{BIC} = \ln \left(\frac{\sum_{i=1}^n \varepsilon_i^2}{n} \right) + \frac{\lambda \ln(n)}{n} \quad (3.5.2)$$

The above two model selection criteria deal with the sum of squared residuals for including additional parameters in the model. Moreover, AIC is a biased estimator of the discrepancy between all candidate models and the true model. To overcome this drawback, “corrected” version of AIC was developed.

(iii) Corrected Aikaiki Information Criteria:

The concept of corrected Aikaiki Information Criteria was originally proposed for linear regression model by Sugiura (1978). Mathematically, BIC is calculated by the equation-

$$\text{AICc} = \ln \left(\frac{\sum_{i=1}^n \varepsilon_i^2}{n} \right) + \frac{2n(\lambda + 1)}{n - \lambda - 2} \quad (3.5.3)$$

3.6 Diagnostic or Model Checking:

After a proposed model is fitted to the data, the next step is checking the model's adequacy or diagnostic. This means that it is necessary to check whether the model is adequate to the data and if there is any evidence of inadequacy then the model should be modified in the next iterative phase. Different techniques can be applied to check the adequacy of the model and they are discussed below:

(i) Diagnostic Checks of the Residuals:

Another important technique of model checking is to look at whether the residuals follow white noise or not. If a time series consists of uncorrelated observations and has constant variance, then the series is said to be white noise. On the other hand, if the residuals violate the white noise assumption, then a new model should be selected by the model

selection criterion. Besides this, another important technique of model checking is to inspect the residuals of ACF and PACF. Moreover, normality of the residuals can be examined from the normal probability plot of the residuals or applying the one sample Kolmogorov-Smirnov test. If the forecast errors are dense or cluster around the straight line, then the residuals are normally distributed and if the errors deviate somewhat from the straight line then the residuals don't follow the normal distribution. Moreover, Kolmogorov Smirnov test assumes that the data is normally distributed and if the p-value of the test statistic is less than 0.05 or 0.01 then the assumption may be rejected and it implies that the data is not normally distributed.

(ii) Portmanteau Lack-of-fit-Test:

Box and Pierce (1970) developed a procedure to assess whether a set of autocorrelations jointly follow white noise criteria. The test statistic is defined as -

$$Q = n \sum_{k=1}^m \hat{\rho}_k^2 \quad (3.6.1)$$

Where, n = number of observations used to fit the model, m = lag length, ρ_k = autocorrelation coefficient.

Here the test statistic Q is approximately distributed as χ^2 distribution with m degrees of freedom. On the other hand, if the model is inappropriate to the data, then the Q statistic's average values will be inflated. To overcome the above difficulty, an approximate “**portmanteau**” test of the hypothesis of the model adequacy was designed to take into account by referring an observed value of the Q statistic to the percentage points of the χ^2 distribution.

3.7 Measures of Forecast Accuracy:

If the residuals follow white noise criterion then the model can be used for forecasting. To choose a final model for forecasting the accuracy of the model must be higher than that of all the competing models (Dutta et al., 2021). The accuracy of the model can be compared using some statistic such as Root Mean Square Error (RMSE), Mean Percentage Error (MPE), Mean Absolute Percentage Error (MAPE), Mean Absolute Deviation (MAD), Mean Squared Error (MSE), Mean Absolute Scale Error (MASE) etc. A model with a minimum of these statistic is considered to be the best for forecasting (Dutta et al., 2021).

The one- step-ahead forecast error is defined as $\varepsilon_t = y_t - \hat{y}_t$. Where ε_t = forecast error at time period t, y_t = actual observations at time period t, \hat{y}_t = forecasted values at time t.

(i) Mean Absolute Error (MAE) and Mean Squared Error (MSE):

The two most commonly used scale-dependent measures are based on the absolute error or squared errors-

$$MAE = \text{mean}|\varepsilon_t| \quad (3.7.1)$$

$$MSE = \text{mean}(\varepsilon_t)^2 \quad (3.7.2)$$

MAE is preferably used to comparing forecast methods on a single series, as it is easy to understand and compute. On the other hand, this approach cannot be used to make comparisons between series as it makes no sense to compare accuracy on different scales.

(ii) Mean Absolute Percentage Error (MAPE):

The MAPE is written as -

$$MAPE = \frac{1}{n} \sum_{t=1}^n \frac{\varepsilon_t}{A_t} \times 100 \quad (3.7.3)$$

In equation (1.5.3), n represents the number of time periods, ε_t stands for forecast error in time period t and A_t denotes actual number of observations at time period t.

Lewis (1982) made the following interpretation of MAPE values as follows: (i) if the value of MAPE is less than 10% then the model performed highly accurate forecasting(ii) if the value of MAPE is between 10%-20% then the forecasting performed of the model is regarded as good(iii) if the value of MAPE is between 21%-50% then it is considered as reasonable forecasting and finally (iv) if the value of MAPE is 51% and above it is considered as inaccurate forecasting.

(iii) Root Mean Square Error (RMSE):

This approach compares the actual rates of changes in time series data and computes the averages forecast errors. It is given by

$$RMSE = \sqrt{\frac{1}{n} (A_t - F_t)^2} \quad (3.7.4)$$

In equation 1.5.4, n is the number of time periods while A_t is the actual number of observations at time period t and F_t represents forecasted values at time t.

(iv) Mean Absolute Scale Error (MASE):

The concept of MASE was proposed by Hyndman and Koehler (2006) which is an applicable measure of forecast accuracy and it is defined as-

$$\text{MASE} = \text{mean} \left(|q_t| \right) \quad (3.7.5)$$

$$\text{Where } q_t = \frac{\varepsilon_t}{\frac{1}{n} \sum_{i=2}^n |y_i - y_{i-1}|}$$

This method can be used to compare forecast accuracy between the series as it is a scale-free measure and it is the only accessible method which can be used in all conditions.

3.8 Box-Jenkins Approach:

The methodology and the theorems propounded by Box and Jenkins (1970) called the Autoregressive Integrated Moving Average (ARIMA) is an advance technique of forecasting requires long time series data. This model decomposes historical data into an Autoregressive (AR) process, where there is a memory of past values, an Integrated (I) process, which accounts for stabilizing or making the data stationary plus a Moving-Average (MA) process, which accounts for previous error terms making it easier to forecast. The original Box-Jenkins modelling technique involved three-stage iterative procedure such as model selection, parameter estimation and model checking. Recent explanations of the process recommend by Makridakis et al., (1998) include a preliminary stage of data preparation and final stage of model application or forecasting.

The autoregressive integrated moving average (ARIMA) model of Box and Jenkins (1970) is given by

$$\phi(B)\nabla^d x_t = \mu + \theta(B)e_t, \quad (3.8.1)$$

Where e_t is the usual white noise process. The general model is denoted by ARIMA (p, d, q). The ordinary autoregressive and moving average components are represented by the following polynomials $\phi(B)$ and $\theta(B)$ of orders p and q, respectively,

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p \quad (3.8.2)$$

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \dots + \theta_q B^q \quad (3.8.3)$$

and the difference components are represented by $\nabla^d = (1-B)^d$.

CHAPTER-IV

ANALYSIS AND INTERPRETATION

4.1. Introduction:

Statistical analysis is the main step in any statistical investigation. After collecting data, data must be scrutinized, edited and tabulated and then a very careful statistical analysis is to be made and finally a report comporting detailed statement to be different stages of the survey should be prepared. In this chapter, required computations are made using the method of described in chapter III, and interpretations are made accordingly.

All the data are analysed and graphics are drawn in R Studio.

4.2 Findings of the Study:

This section is divided into two parts. First part involves testing the Model and the later part discussed the validation of the model.

4.2.1 Testing Part:

4.2.1.1 Data Preparation and Transformation:

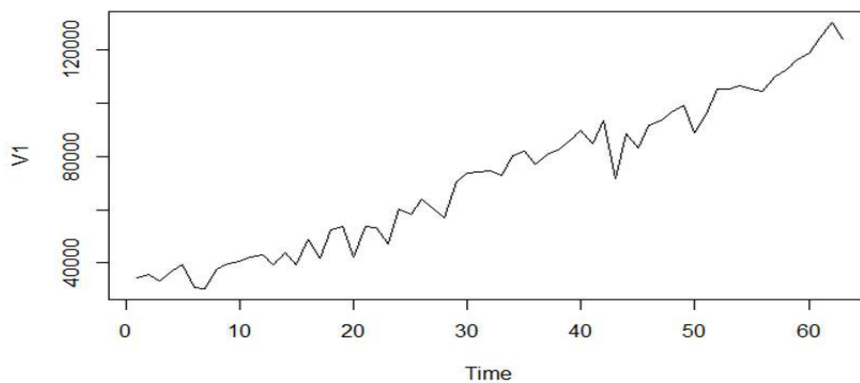


Fig4.2.1. Original plot of Rice Production data in India

Fig. 4.2.1 shows the yearly Rice Production Data in India from **1960-2022**. From the fig.4.2.1, it is observed that the time plot exhibits a significant upward trend. Next, the whole data set divided into two parts: testing (1960-2012) and validation (2013-2022). First, the researchers develop the ARIMA model for testing part and compare the forecasted values from the model with validation part. If the model fits well and fulfills the all assumptions,

then refit the model for the whole data set i.e. 1960-2022 and forecast the amount of yearly production of Rice for the upcoming 10 years.

Next, Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) of testing data set in Fig. 4.2.2 reveal the significant trend in the data set. But, according to B-J methodology, we must ensure that the time series being analyzed is stationary before fit the ARIMA model.

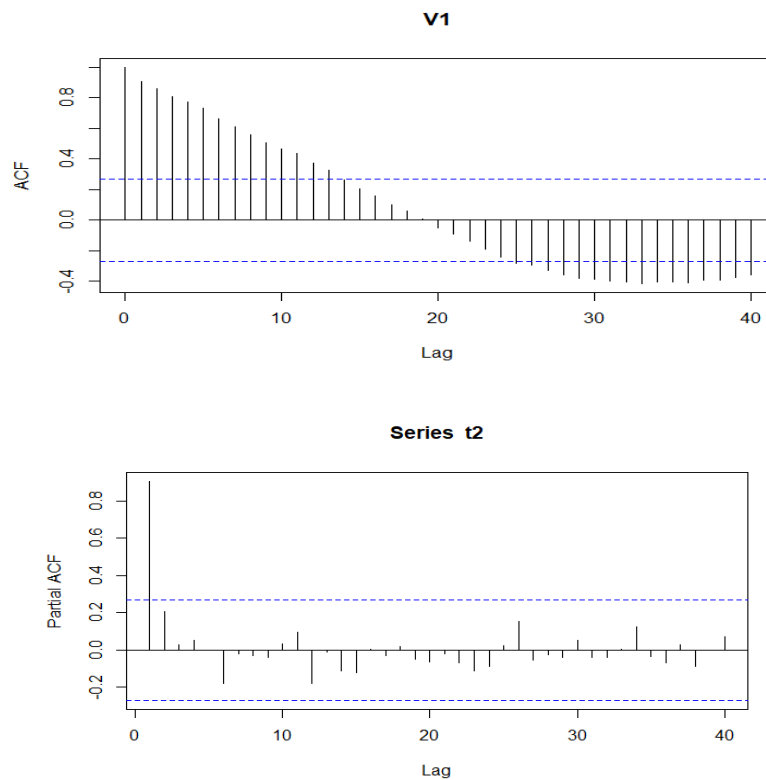


Fig 4.2.2: ACF and PACF of testing data set

Therefore, in order to obtain a stationary series, the researcher decides to first take first differences of data to remove trend from the series. Fig.4.2.3 consists of plotting first difference of the data.

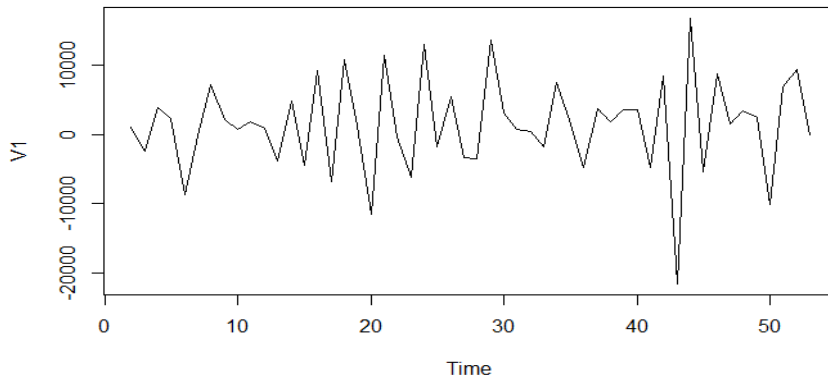


Fig 4.2.3: First difference of testing data

Fig.4.2.3, showed a slide increasing pattern of the difference data. Also, by applying augmented Dickey-Fuller test to the first differenced data it is found that $p\text{-value} < 0.05$, so we may accept the null hypothesis i.e. the first differenced series is stationary. Further, ACF and PACF of the first differenced of the testing data are plotted in Fig. 4.2.4

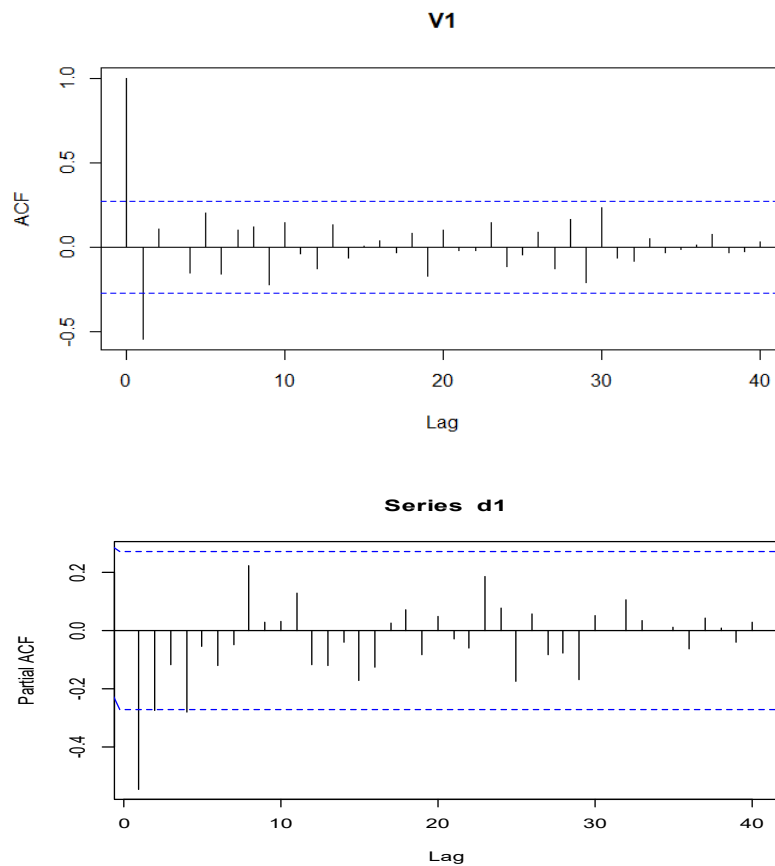


Fig 4.2.4: ACF and PACF of first difference data

From fig 4.2.4, it is observed that there is a single significant spike at lag 1 in the ACF process indicating Moving Average (MA) Process of order 1 and two significant spikes at lag one and two in PACF indicating Auto Regressive Process of order 2. So, it might tentatively conclude that the time series values are described by ARIMA model of order $p=2, d=1, q=1$. i.e. ARIMA (2,1,1) model is suggested by the data. Next, to estimate the parameters of the model the researcher consider different grouping of (p, q) with the difference of order ($d=1$) which is taken in advance.

4.2.3 Model identification and estimation of parameters:

The estimated parameters by maximum likelihood method obtained from R software are as follows:

Table 4.2.1: Parameter Estimation of the proposed ARIMA Models:

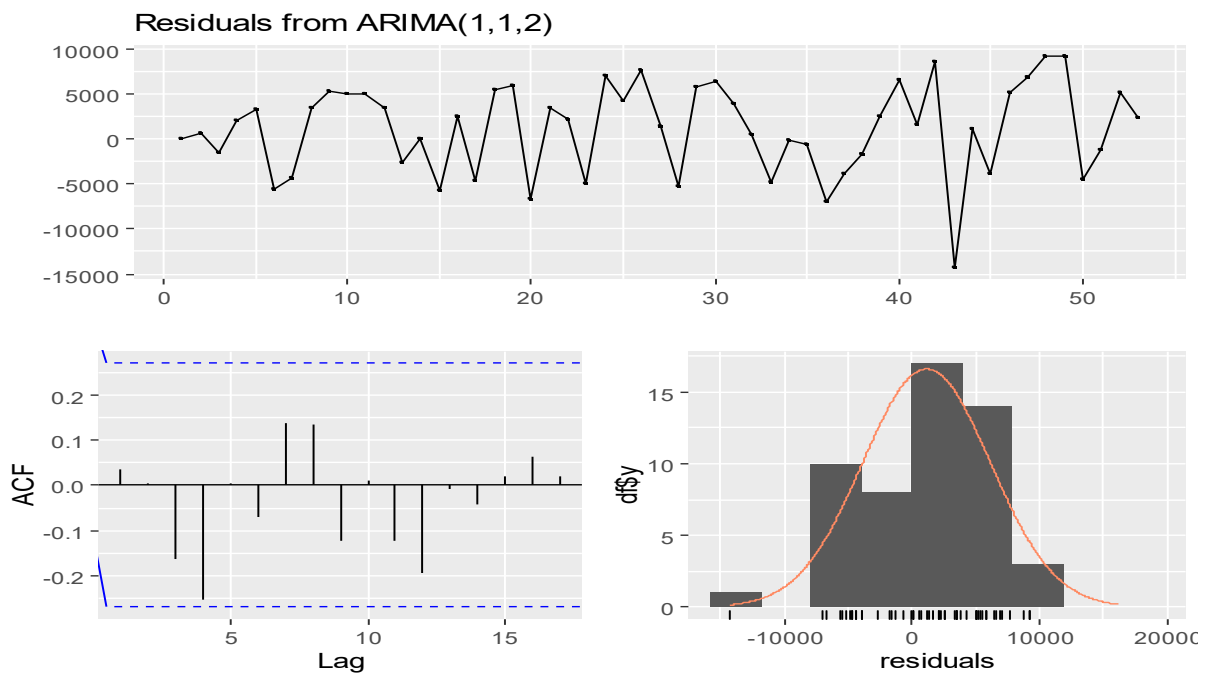
| Models | Parameters | Estimates | S.E. | Z -Value | p-value |
|---------------|------------|-----------|----------|----------|--------------|
| ARIMA (2,1,1) | ϕ_1 | 1.14630 | 0.36841 | -3.1115 | 0.001862** |
| | ϕ_2 | 0.41733 | 0.17664 | -2.3625 | 0.018150* |
| | θ_1 | 0.62462 | 0.38260 | 1.6326 | 0.10255 |
| ARIMA (2,1,0) | ϕ_1 | -0.5408 | 0.1369 | -3.9516 | 7.762e-05** |
| | ϕ_2 | -0.1306 | 0.1380 | -0.9459 | 0.3442 |
| ARIMA(1,1,0) | ϕ_1 | -0.4764 | 0.1196 | -3.9852 | 6.743e-05** |
| ARIMA(0,1,1) | θ_1 | -0.4426 | 0.0992 | -4.4633 | 8.072e-06*** |
| ARIMA(1,1,2) | ϕ_1 | 0.88542 | 0.070146 | 12.533 | <2.2e-16*** |
| | θ_1 | -1.810723 | 0.089949 | -20.131 | <2.2e-16*** |
| | θ_2 | 0.99979 | 0.095883 | 10.429 | <2.2e-16*** |
| ARIMA(1,1,1) | ϕ_1 | -0.3169 | 0.2366 | -2.5431 | <2.2e-15*** |
| | θ_1 | -0.2148 | 0.2384 | -8.9765 | <1.2e-13*** |

Table 4.2.2: Accuracy Measurement of the Suggested Models:

| Models | AIC | AICC | BIC | RMSE | MAPE | MASE |
|---------------------|----------------|----------------|----------------|----------|-----------------|------------------|
| ARIMA (2,1,1) | 1060.33 | 1061.18 | 1068.13 | 5925.045 | 8.245005 | 0.93111 |
| ARIMA (2,1,0) | 1058.91 | 1059.41 | 1064.76 | 5960.001 | 8.346319 | 0.9478081 |
| ARIMA(1,1,2) | 1050.62 | 1051.47 | 1058.43 | 5091.612 | 7.285887 | 0.8143176 |
| ARIMA(1,1,1) | 1059.08 | 1059.58 | 1064.93 | 6013.101 | 8.431175 | 0.9492268 |
| ARIMA(1,1,0) | 1057.79 | 1058.04 | 1061.7 | 6013.101 | 8.431175 | 0.9492268 |
| ARIMA(0,1,1) | 1058.7 | 1058.95 | 1062.61 | 6068.246 | 8.428838 | 0.9644695 |

Table 4.2.2 shows the accuracy measurement of the suggested models. Here, Aikaiki Information Criterion (AIC), Aikaiki Information Criterion Corrected (AIC_C), Bayesian Information Criterion(BIC), Root mean square Error (RMSE), Mean Absolute Percente Error (MAPE), Mean Absolute Scaled Error (MASE) are presented against the models. Comparing all these measurements, it is observed that ARIMA(1,1,2) model has lowest AIC, AIC_C , BIC, RMSE, MAPE and MASE.

Next, the reseacher check the adequacy of the fitted models. The following figures demonstrate the plot of standardized residuals, ACF of residuals, Normal Q-Q Plot of standardized residuals and p-values for Ljung –Box statistics for each of the fitted models.



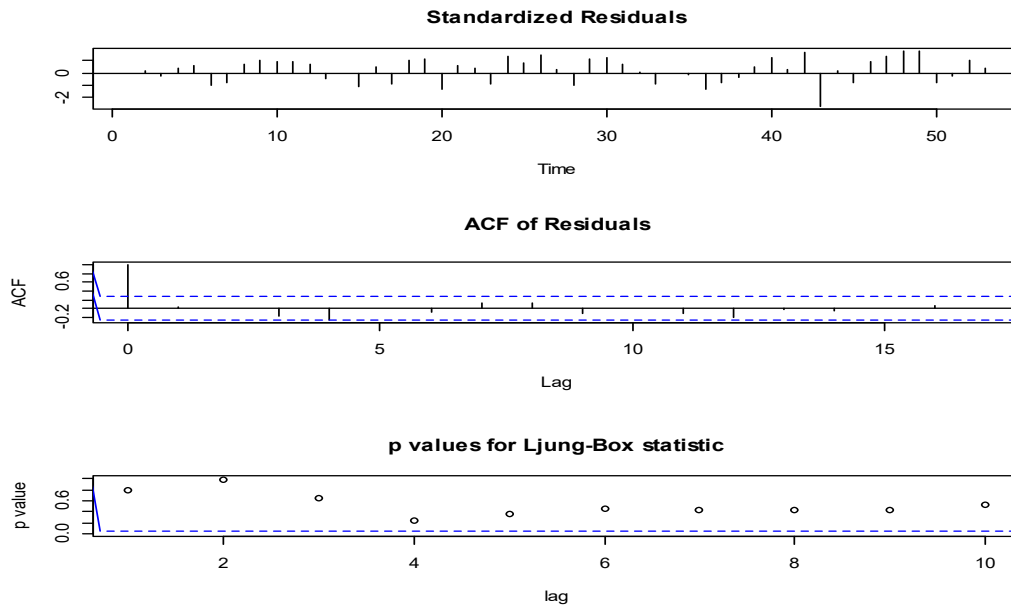


Fig 4.2.5: Plot of standardized residuals, the ACF of the residuals, and the p-values of the Q-statistic for ARIMA (1, 1, 2) Model.

From fig 4.2.5, it is observed that plot of the standardized residuals shows no obvious pattern and trend and looks like an independent and identical distribution. From standardized plot of residuals, it is also observed that residuals are lie within ± 3 . Here, none of the autocorrelation is individually significant nor the Ljung-Box-Pierce Q Statistics are statistically significant. Here the value of Q statistics is 9.1269 and p-value is 0.5201, so the null hypothesis of independence of residuals cannot be rejected. Using the white noise test from (the normwhn.test package) in R, perform a univariate test for white noise, it is obtained that p-value is 0.2276 which means that residuals series is white noise (with mean 0 and variance σ^2). The plot of the ACF of the residuals of the diagnostics shows no evidence of significant correlation in the residuals. Most of the standardized residuals are located on the straight line except few outliers deviating from the normality. And lastly the bottom part of the diagnostics is the time plot of the Ljung- Box statistics which shows that the statistic is not significant at any positive lags.

4.2.4 Model Application:

The maximum likelihood estimates of the parameter of the model ARIMA (1,1,2)Model already depicted in Table 4.2.1.

Now, the forecasted number of yearly production of rice in India from the fitted model presented in the following table.

Table 4.2.3 Forecasted number of yearly production of Rice in India from ARIMA (1,1,2) Model for upcoming 10 years:

| Year | Point forecast | 95% Confidence Interval | |
|------|----------------|-------------------------|---------|
| | | Lower | upper |
| 2013 | 106075 | 99166 | 112984 |
| 2014 | 109089 | 102146 | 116034 |
| 2015 | 111745 | 104543 | 118946 |
| 2016 | 114083 | 106294 | 121871 |
| 2017 | 116142 | 107433 | 124849 |
| 2018 | 117954 | 108051 | 127857 |
| 2019 | 119549 | 108252 | 130847 |
| 2020 | 120955 | 108129 | 133782 |
| 2021 | 122193 | 107754 | 1366632 |
| 2022 | 1123283 | 107182 | 139383 |

4.2.5 Validation Part:

Now, to check the validity of the fitted model, the actual observations are shown against the predicted values from 2013-2022 (10 years) in the following table along with 95% confidence interval.

Table 4.2.4 Actual and forecasted observations from ARIMA(1,1,2) Model

| Year | actual observation | point forecast | 95% CI | |
|------|--------------------|----------------|--------|--------|
| | | | Lower | upper |
| 2013 | 105725 | 101726 | 95508 | 116641 |
| 2014 | 114590 | 105199 | 98469 | 119710 |
| 2015 | 118239 | 108671 | 100731 | 122758 |
| 2016 | 126896 | 112143 | 10172 | 125994 |
| 2017 | 133938 | 115616 | 102823 | 129460 |
| 2018 | 136834 | 119088 | 102808 | 133099 |
| 2019 | 139091 | 122560 | 102272 | 136828 |
| 2020 | 137423 | 126033 | 101339 | 140571 |
| 2021 | 141526 | 129505 | 100111 | 144275 |
| 2022 | 148707 | 132977 | 98659 | 147906 |

From the above table, it is observed that the amount of yearly rice production in India from 2013-2022 is almost equal and exact pattern with the actual data. It is also observed that actual observations are also within the prediction limit (95%). Therefore, it may be concluded

that proposed model would be good fitted to the yearly rice production in India from 1960-2022.

Further, ARIMA (1,1,2) Model is used to forecast the yearly rice production data in India for the upcoming 10 years.

Table 4.2.4 Forecasted amount of Rice Production in India from ARIMA (1,1,2) Model for upcoming 10 years:

| Year | Point forecast | 95% confidence Interval | |
|------|----------------|-------------------------|--------|
| | | Lower | Upper |
| 2023 | 128186 | 117594 | 138778 |
| 2024 | 130244 | 118982 | 141505 |
| 2025 | 132288 | 120284 | 144294 |
| 2026 | 134321 | 121504 | 147139 |
| 2027 | 136342 | 122648 | 150037 |
| 2028 | 138351 | 123720 | 152981 |
| 2029 | 140347 | 124726 | 155969 |
| 2030 | 142332 | 125669 | 158996 |
| 2031 | 144305 | 126553 | 162058 |
| 2032 | 146266 | 127381 | 165152 |

CHAPTER-V

CONCLUSION

In this study, Box-Jenkins (ARIMA) methodology is used to forecast the yearly Rice Production Data in India. The estimation and diagnostic analysis results revealed that the models are adequately fitted to the historical data. The present study reveals that ARIMA (1, 1, 2) Model would be good fitted to the yearly rice production data in India. Moreover, ARIMA (1,1,2) Model gives the 10 years forecasted Rice production. The forecasted Rice productivity revealed an increasing pattern for the upcoming 10 years and it has been estimated as **146266** MT in the year 2032. This may be a good indication to us.

As such our predicted Rice productivity pattern may add to the existing knowledge of agricultural productivity in India, where the population increases day by day and food is a major concern. These projections can play a vital role to deal with future food security measures and planning for policy makers in India.

REFERENCES:

1. Biswas, R. and Bhattacharyya, B., (2013): “ARIMA Modeling to Forecast Area and Production of Rice In West Bengal”, *Journal of Crop and Weed*, 9(2),pp. 26-31.
2. Hazarika J., Pathak B. and Patowary A. N. (2017): “Studying Monthly Rainfall over Dibrugarh: Use of SARIMA Approach”, *MAUSAM*, 68(2), pp.349-356.
3. Monfared A.B., Soori H., Mehrabi Y., Hatami H. and Delpisheh A. (2013): “Prediction of Fatal Road Traffic Crashes in Iran using the Box-Jenkins Time Series Model”, *Journal of Asian Scientific Research*, 3(4), pp. 425-430.
4. Prabakaran K., and Sivapragasam C., (2014): “Forecasting Areas and Production of Rice in India using ARIMA Model”, *International Journal of Farm Sciences*, 4(1), pp. 99-106.
5. Ramakrishna G., and Vijayakumari R., (2017): “ARIMA model for Forecasting of Rice Production in India by using SAS”, *International Journal of Applied Mathematics and Statistical Science*, 6(4), pp. 67-72.
6. Sahu P.K., Mishra P., Dhekale B.S., K.P. Vishwajith and Padmanaban K., (2015) : “ Modeling and Forecasting of Area, Production, Yield and Total seeds of Rice and Wheat in SAARC Countries and the World towards Food security”, 3(1),pp. 34-48.
