Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

Production Analysis and Prediction of TOP using ARIMA Model

(PAPTOP)

Darshan Halliyavar[1], Shwetha M P[1], Sharanabasavaraj[1], Puneeth S[1] and Shivaprasad Ashok Chikop[2]
[1]Department of Electronics and Communication Engineering, Dayananda Sagar University, Bangalore, Karnataka, India
[2]Department of Computer Science and Engineering, Dayananda Sagar University, Bangalore, Karnataka, India

***Abstract*—Tomato, Onion and Potato are the staple foods that form the essential part of meal in the contemporary world. Lack of co-ordination between the demand, supply and production induces price fluctuation in the developing country like India. In order to overcome this problem a synchronous model should be in place to reduce the price fluctuation. One of the solutions for the mentioned problem, is designed in the work carried out. ARIMA model has been utilized in the work to come up with the best model for predicting the amount of production based on the previous data. The work provides the best AR, MA and ARIMA model and also the prediction for 15 years based on the previous available data. In conclusion, the work provides the ARIMA model for tomato, onion and potato.**

Keywords*—ARIMA, Time Serires Analysis, TOP, Agriculture, Data Analytics.

I. INTRODUCTION

Agriculture plays an important role in Indian economy. It contributes to 17% of total GDP of India and provides employment over 60% of the population [1]. The success of ancient to modern civilization throughout the world comes from the cultivation of crops in different geographic conditions. The cultivation of crops depends upon versatile factors and these factors should be taken care such that studying these different aspects helps to produce a maximum yield for the farmers.

Time Series Analysis is a type of a statistical analysis, which deals with the time series data, and its analysis for regular period of intervals [2]. It is mainly used for predicting the future values based on past observed values. The main objective of time series analysis is to develop the mathematical models which provides sample data and in order to provide a statistical setting for describing the character of data that seemingly fluctuate in a random fashion over time, we assume a time series can be defined as a collection of random variables indexed according to the order they are obtained in time.

It is typically to show an example for time arrangement graphically by plotting the estimations of the random factors on the vertical pivot, or ordinate, with the time scale as the abscissa. It is normally helpful to associate the qualities at nearby time spans to recreate outwardly some unique theoretical consistent time arrangement that may have delivered these qualities as a discrete sample.

Some of the works carried out using the time series analysis in various domains are mentioned: Forecasting Wi-Fi data network traffic is determined by using time series analysis, Box-Jenkins methodology and ARIMA model [3]. Sample data of correlated structure is taken and different time series analysis methods are observed for data traffic detection. Mean square, standard deviation and correlation co-efficient are calculated to know the relational difference between two values [3]. By using Time series approach, demand in food industry is modeled and forecasted. ARIMA model and Box-Jenkins methodology used to identify four performances criteria such as Akaike criterion, Schwarz Bayesian criterion, maximum likelihood, and standard error. If there exist any nonlinearity in the obtained data set then ANN (Artificial neural network) technique is imposed. The obtained results are again compared with available historical data to forecast the further demand in food industry [4]. The prediction of Sugarcane production in India is done by using Time series analysis, ARIMA model and Box-Jenkins methodology. The work has carried out in forecasting the production of sugarcane in five leading years. Model identification has done to check and identify the variable to be forecasted is stationary in time series. Augmented Dickey-Fuller (ADF) Test is carried for stationary test. Autocorrelations (ACF) and Partial Autocorrelations (PACF) are used for first order differenced series [5]. With the conclusion ARIMA model plays very important role, in order to forecast future successive values in the time series to the work. While exponential smoothing models are based on a description of the trend and seasonality in the data [2], ARIMA models aim to describe the autocorrelations in the data [2]. In this regard, Autoregressive Integrated Moving Average (ARIMA) model helps farmers to choose the future scope of agricultural crops in the best way by using the past values and the present results. Exponential smoothing and ARIMA models are the two methods used for different approaches in time series forecasting, and further provides complementary approaches

This work focuses on forecasting the fluctuation in the production of TOP. The word TOP refers to "Tomato, Onion, Potato" which are one of the most essential vegetables which are grown all over the world [6]. These vegetables are used very prominently in daily life and are grown in all time irrespective of seasons [6]. This makes space for volatile supply of TOP in market. The production of these crops varies and due to which a specific model need to be set that provide information regarding stipulated time frame for crop cultivation. This provides information for setting up minimum selling price to farmers. Due to fluctuation in their prices the Indian Government has made '**Operation Greens'** which is a project approved by the Ministry of Food Processing

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

Industries with the target to stabilize the supply of tomato, onion and potato crops (TOP crops) in India [7], as well as to ensure their availability around the country throughout the year. These crops can be grown such that they are not affected by MSP rates [11].

II. MATERIALS AND METHODS

As mentioned in the introduction section TOP are considered as the essential part of the meal. The demand for TOP will be for throughout the year. The work was carried out in two phases. The first phase includes data acquisition and data structuring. The second phase comprises of data analysis. The steps carried out in the work are summarized in Fig 1. The acquired data was not present in the required format, thus it was structured in the required format. The ARIMA method was utilized for analysis of the previous data to predict the values. The values might help to take better decision-making for sowing, budgeting, expected production and insights into import/export of TOP. The ARIMA model is integration of two independent models Auto regressive (AR) and Moving Average (MA). This model does not assume any specific pattern on the time series data that is utilized for the prediction. The ARIMA model is based on the following building blocks: (i) Model identification (ii) Parameter estimation (iii) Model validation

(i) Model identification: The model identification specifies the orders (p, d, q) of two components of ARIMA model. Typically, order's function is to determine the signal either stationary or non-stationary. If the signal is non-stationary the signal has to be converted to stationary.

(ii) ***Parameter estimation:*** The model cannot be structured until we obtain the stationary signal. The non-stationary signal can be converted to stationary signal by determining the d-value. The optimal selection of the d-value is a challenge in itself. To address this issue it is advised to always start with minimum value. If a higher value is selected then it might lead to larger standard deviation. The optimal d value that provides with zero mean and standard deviation of the signal is considered as the stationary signal. In order to obtain stationary signal ADF test was carried out.

(iii) ***Model validation:*** In order to validate the model ACF and PACF were carried on the obtained stationary signal. In the final step based on the output of ACF and PACF the optimal model of ARIMA was selected. The obtained model was used for the prediction of TOP crop. The implementation of the work was carried out using Python Jupiter Notebook. The Time series flow was inspired by course taken from Edureka. [12].

The data for the work was collected from National Horticultural Research and Development Foundation. The collected data was from January 2013 to December 2019 sampled monthly. The data represents the annual production of TOP in terms of quintal all over India.

***Tomato*:** The suitable mathematical model in case of tomato was: AR (2, 1, 1), MA (0, 1, 2), ARIMA (2, 1, 1). The obtained result on the data analysis of tomato vegetable is as depicted in the fig 2.

Onion: The suitable mathematical model in case of onion was: AR(2, 1, 1), MA(0, 1, 2), ARIMA(2, 1, 2). The obtained result on the data analysis of onion vegetable is as depicted in the fig 3.

Potato: The suitable mathematical model in case of potato was: AR(2, 1, 1), MA(0, 1, 2), ARIMA(2, 1, 1). The obtained result on the data analysis of potato vegetable is as depicted in the fig 4.

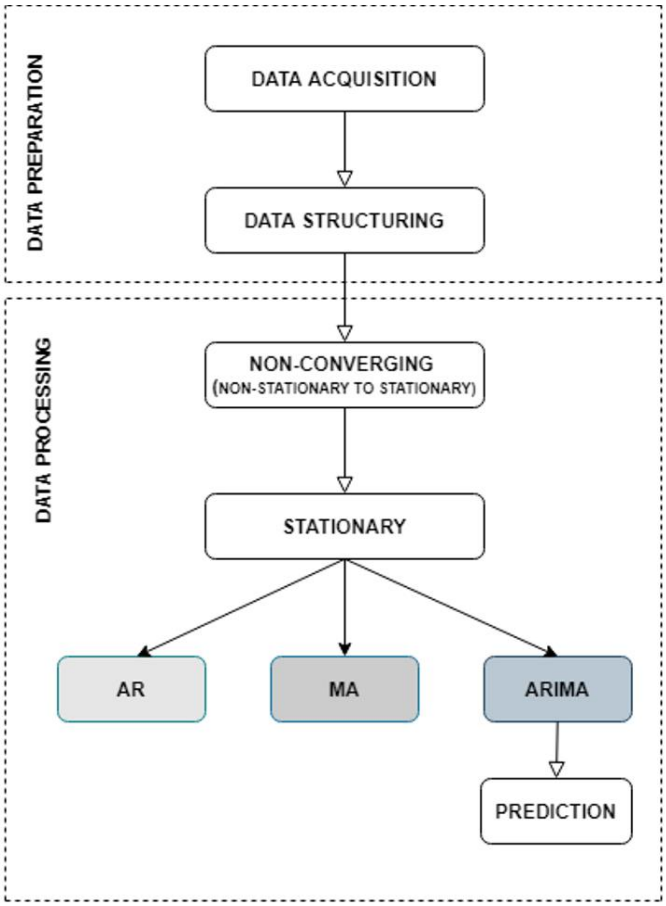

Fig 1: Summarizes the work flow with in the designed model

III. RESULTS

Tomato: Fig 2 depicts the intermediate and prediction curve. Fig 2a shows the monthly prediction curve from the year 2013 to 2019 along X-axis and the corresponding production along Y-axis. The corresponding stationary curve in log scale is demonstrated in fig 2b. The statistical curve obtained for AR, MA, ARIMA model are represented in Fig 2(c, d, e) respectively. The blue colored curve is the statistical and red is its corresponding mean. The prediction obtained through ARIMA model for Tomato is depicted in Fig 2f.

The statistical parameters after Dickey-Fuller Test for Non-Stationary and Stationary are mentioned below:

Non-Stationary

Test Statistics	3.342951
p-value	0.813064
Lags Used	12.000000
Number of Observations Used	71.000000
Critical Value (1%)	3.526005
Critical Value (5%)	2.903200
Critical Value (10%)	2.588995

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

Stationary

Test Statistics	-1.611182
p-value	0.277355
Lags Used	12.000000
Number of Observations Used	65.000000
Critical Value (1%)	-4.473135
Critical Value (5%)	-3.289881
Critical Value (10%)	-2.772382

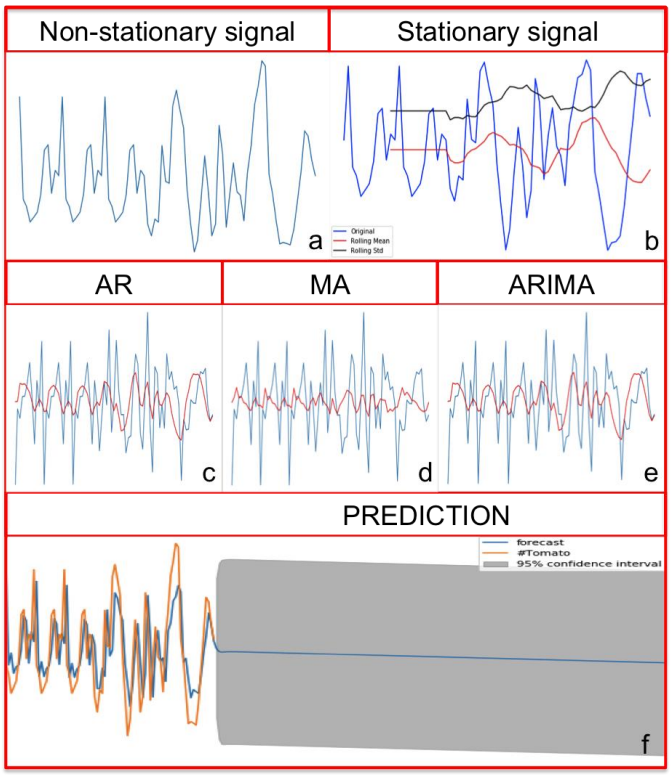

Fig 2: (a) Shows the non-stationary curve from the input data of tomato, (b) corresponding stationary signal at log scale, (c, d, e) the curve obtained from AR, MA, ARIMA model at log scale (blue color is the actual curve and red depicts the respective mean value), (f) orange curve is the actual signal, blue is the corresponding forecast (grey region in the confidence region of 95%)

Onion: Fig 3 shows the intermediate and prediction curve. Fig 3a depicts the monthly prediction curve from the year 2013 to 2019 along X-axis and the corresponding production along Y-axis. The respective stationary curve in log scale is demonstrated in fig 3b. The statistical curve obtained for AR, MA, ARIMA model are represented in Fig 3(c, d, e) respectively. The blue colored curve is the statistical and red is its corresponding mean. The prediction obtained through ARIMA model for Onion is depicted in Fig 3f.

The statistical parameters after Dickey-Fuller Test for Non-Stationary and Stationary are mentioned below:

Non-Stationary

Test Statistics	9.469114
p-value	4.152202
Lags Used	5.000000
Number of Observations Used	8.000000
Critical Value (1%)	3.511712
Critical Value (5%)	2.897048
Critical Value (10%)	2.585713

Stationary

TestStatistics	-0.000000
p-value	0.958532
#Lags Used	5.000000
Number of Observations Used	8.000000
Critical Value (1%)	-7.355441
Critical Value (5%)	-4.474365
Critical Value (10%)	-3.126933

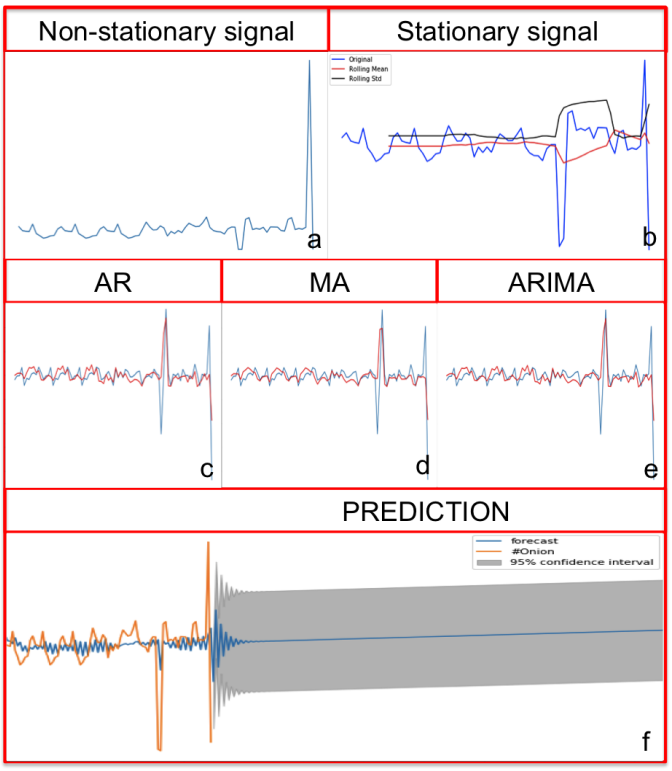

Fig 3: (a) Shows the non-stationary curve from the input data of onion, (b) corresponding stationary signal at log scale, (c, d, e) the curve obtained from AR, MA, ARIMA model at log scale (blue color is the actual curve and red depicts the respective mean value), (f) orange curve is the actual signal, blue is the corresponding forecast (grey region in the confidence region of 95%)

Potato: Fig 4 depicts the intermediate and prediction curve. Fig 4a shows the monthly prediction curve from the year 2013 to 2019 along X-axis and the corresponding production along Y-axis. The corresponding stationary curve in log scale is demonstrated in fig 4b. The statistical curve obtained for AR, MA, ARIMA model are represented in Fig 4(c, d, e) respectively. The blue colored curve is the statistical and red is its corresponding mean. The prediction obtained through ARIMA model for Potato is depicted in Fig 4f.

The statistical parameters after Dickey-Fuller Test for Non-Stationary and Stationary are mentioned below:

Non-Stationary

Test Statistics	1.606762
p-value	0.480196
#Lags Used	12.000000
Number of Observations Used	71.000000
Critical Value (1%)	3.526005
Critical Value (5%)	2.903200
Critical Value (10%)	2.588995

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

Stationary	
Test Statistics	-4.458257
p-value	0.000234
#Lags Used	12.000000
Number of Observations Used	65.000000
Critical Value (1%)	-5.354256
Critical Value (5%)	-3.646238
Critical Value (10%)	-2.901198

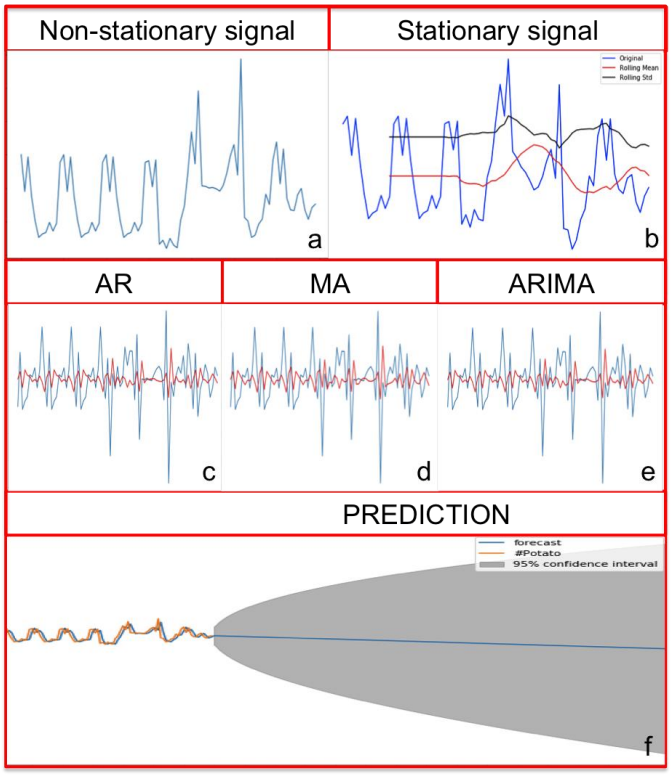

Fig 4: (a) Shows the non-stationary curve from the input data of potato, (b) corresponding stationary signal at log scale, (c, d, e) the curve obtained from AR, MA, ARIMA model at log scale (blue color is the actual curve and red depicts the respective mean value), (f) orange curve is the actual signal, blue is the corresponding forecast (grey region in the confidence region of 95%)

IV. DISCUSSION AND CONCLUSION

The prediction curve obtained for tomato depicts a downward trend as shown in fig 2f. This provides the information that the production of tomato may go down. Hence the precautionary measures have to be taken care for the corresponding year to maintain the demand and supply chain. The downward trend also tries to provide inputs in decision making for importing the tomato into the country. This in turn helps to prevent fluctuation impinging the market. The statistical parameters as mentioned in the results section for tomato can be improved. The p-value obtained can be improved by incorporating more data.

The prediction curve obtained for onion shows upward trend as shown in fig 3f. This provides the information that the production of onion may rise. Measures can be taken for further improvement of production of onion within the country. This gives insight into the quantity of onion to be imported/exported based on the demand/supply for the corresponding year. The statistical parameters obtained in case of onion are not promising hence the accuracy of prediction can be improved by employing more historical data.

The prediction curve obtained in case of potato is optimal as compared to other prediction curves this is due to the ARIMA model employed incase of potato is optimal. The curve depicts the improved production of potato in the upcoming year as shown in the fig 4f. The accuracy of model prediction can be improved further by incorporating more data. The statistical parameters as mentioned in the results section are satisfactory.

The developed model do not provide insights on the other agricultural related parameters such as weather forecast, soil fertility and approximate time for sowing for improving the production of TOP. Incorporating the above mentioned parameters in the model can improve the region wise prediction and also production of TOP. The region wise prediction is necessary as India has 16 kinds of climate zones [10]. If prediction of the TOP is tailored region wise, it will help reduce the need of transportation from one region to other. It will stimulate the region wise production of TOP. This will help in maintaining the cost fluctuation and avoid transportation cost.

The data available from the source was limited and due to which the prediction has a lower accuracy. So the focus of future work is to improve the accuracy by incorporating more data. Along with improving the accuracy of the model a framework will be designed to provide more insights for optimizing the production of TOP. The outcome of the framework will be to provide information related to all the parameters [5].

In conclusion, the developed model can be utilized as a tool for analysis of budget allocation for importing the necessary crop in case of shortage. This in turn help to provide insight on the abrupt fluctuations in the market specifically for the onion. The insights might be used for better decision making in order to overcome the market fluctuation. On the other hand the data analysis from the forecast can also be utilized to optimize the production of the TOP and avoid the upcoming shortage of the crop at a particular time of the year.

References

[1] Arjun, Kekane Maruti. "Indian agriculture-status, importance and role in Indian economy." International Journal of Agriculture and Food Science Technology 4.4 (2013): 343-346.

[2] Shumway, Robert H., and David S. Stoffer. Time series analysis and its applications: with R examples. Springer, 2017

[3] Cesar Augusto Hernández Suarez1, Octavio José Salcedo Parra2 y Andrés Escobar Díaz3, An ARIMA model for forecasting Wi-Fi data network traffic values Modelo ARIMA para pronosticar valores de tráfico en una red de datos Wi-Fi, REVISTA INGENIERÍA E INVESTIGACIÓN VOL. 29 No. 2, AGOSTO DE 2009 (65-69)

[4] Jamal Fattah1, Latifa Ezzine1, Zineb Aman2 , Haj El Moussami2, and Abdeslam Lachhab1, Forecasting of demand using ARIMA model, International Journal of Engineering Business Management Volume 10: 1–9 ª The Author(s) 2018.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 07, July-2020

[5] Kumar Manoj, Anand Madhu, Application Of Time Series Arima Forecasting Model For Predicting Sugarcane Production In India, Studies In Business And Economics.

[6] H M, Swamy. (2020). Green farming paper on operations green.

[7] https://mofpi.nic.in/Schemes/operation-greens

[8] Rahman, Sk Al Zaminur, Kaushik Chandra Mitra, and SM Mohidul Islam. "Soil classification using machine learning methods and crop suggestion based on soil series." In 2018 21st International Conference of Computer and Information Technology (ICCIT), pp. 1-4. IEEE, 2018.

[9] Fattah, Jamal, Latifa Ezzine, Zineb Aman, Haj El Moussami, and Abdeslam Lachhab. "Forecasting of demand using ARIMA model." International Journal of Engineering Business Management 10 (2018): 1847979018808673.

[10] Kumar, Ajay, Mokbul Morshed Ahmad, and Pritee Sharma. "Influence of climatic and non-climatic factors on sustainable food security in India: a statistical investigation." International Journal of Sustainable Agricultural Management and Informatics 3, no. 1 (2017): 1-30.

[11] Pandey, R. K. "The Analysis of Demand for Foodgrain." Indian Journal of Agricultural Economics 28, no. 902-2018-2203 (1973): 49-55.

[12] https://www.edureka.co/data-science-python-certification-course