ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2024.0473

YIELD PREDICTION SYSTEM – A SYSTEMATIC APPROACH FOR PREDICTING AGRICULTURAL COMMODITY

K. Vikranth and K. Krishna Prasad
Institute of Computer Science and Information Science, Srinivas University, India

Abstract

In India, the agriculture is main profession for more than sixty percent of the population. The stakeholders of agriculture in India, facing plenty of problems that leads the people of the country to shift their profession and lets them migrate towards urban area. So, need of implementing technology in agriculture is must in future days because as population is increasing in exponential form as the result huge requirement of food and agricultural product. The data analytics will play a significant role in agricultural dataset for implementing prediction and recommendation system in the sector. Yield is one of the factors to be considered in the agriculture that determines the wellness and prosperity of the farmer. In this paper deals with prediction system to predict yield of areca nut product in Puttur taluk, Dakshin Kannda District, Karnataka state in India. The time series data analytics model known as Auto Regressive Integrated Moving Average (ARIMA) model is used for yield prediction system. The research is mainly focused on forecasting of areca nut production for next four or five years in Puttur taluk. It compares various ARIMA models with performance criteria and selects best model for prediction purpose. The diagnostic check is carried out to test the system performance After the prediction, then the actual values and predicted values are compared and presented in the form graphical representation.

Keywords:

ARIMA, Prediction System, Smart Agriculture, Areca Nut

1. INTRODUCTION

The country like India, has significant contribution by agriculture sector to the growth of Indian economy. In entire world the agriculture in India placed second rank and above fifteen percent of the Gross Domestic Product (GDP) was contributed by agriculture sector [1]. The past and present practices used in agriculture trails the sector in all the aspects that fail to meet the worlds food requirement. Here the need of implementing technologies in agriculture in the name of smart agriculture will be the solution for future requirement. The recent emerging technologies like IoT, Artificial Intelligence (AI), Wireless sensor networks (WSN) and data analytics has tremendous impact on smart agricultural system that transforms current agricultural practice in to smart agriculture that solves many problems faced by the farmers [2]. Agriculture has wide area involves multiple aspects such as soil, temperature, water, humidity, animal, live stokes, market price, machinery, etc. There is a need for implementing technologies in every area of the agriculture to transform in to modernization and impart changes in the sector. This paper focused on predictive analytics for predicting useful parameters in the yield of the agricultural commodity. The predictive modelling involved with statistical approaches such as linear and logistic regression to understand trends and predict upcoming parameters, and data mining concept and approach to provide insight and forecasts [3]. The different data analytic techniques are existing and applied to analyze the dataset based on the complexity of the data. The time series data analytic technique is used in this research because of the nature of the dataset. The time series data analytic approach called as Auto Regressive Integrated Moving Average (ARIMA) is the time series approach is well taken to forecast the yield of the areca nut product especially in Puttur Taluk of South Canara district of Karnataka State. The proposed research also has coined its contributions to improve the lifestyle of all stake holders of Areca nut product in Puttur taluk by allowing to take appropriate decision in their agricultural profession. The tremendous popularity of big data and data analytics is today setting a benchmark technology in all most all the sectors like business, agriculture, education, tourism, medical etc. The data analytics is mainly used in recommendation system and prediction system. The research paper is also based on data analytics which examines past fifteen years yield data of areca nut and tries to forecast the next four or five years of yield. The outcome of the research will help farmers and stake holders of areca nut to take appropriate decision and it minimizes the uncertainty problem in yield of areca nut faced by farmers.

2. OBJECTIVES

- To study the different stages involved in data analysis process and significance of each stage for achieving accuracy in result.
- The collected data is examined with different order of ARIMA for the optimum result towards yield prediction.
- To forecast next 4 years of areca nut production by using predictive analytics approach known as ARIMA model.

3. DATA COLLECTION AND METHODOLOGY

The experimental technique is applied in this study, which is concerned with values, comparisons, and behaviour. Year-wise yield data of areca nut of puttur taluk, D.K district in Karnataka is taken from the Open Governmental Data (OGD) platform India's official website from the year 1997 to 2021 [4].

Table.1. The structure of areca nut yield dataset.

Attributes	Data type	Description
Year	Date	Year in the form of YYYY-YY
Area (in Hectare)	Number	Total Area of Taluk Arecanut cultivated
Production (Tonnes)	Number	Total production in Puttur Taluk.

Yield (Tonnes)	Number	Yeild in terms of tonnes per Hectare

The dataset has four attributes, and all four attributes are numerical attributes. It covers the information such as year, Area in terms of Hectare, Production in terms of tonnes, and yield in form of tonnes per hectare. This data set covers year wise yield of the arecanut in terms of tonnes per hectare in Puttur Taluk of D.K district. The below table shows the structure of the dataset.

The research uses time series approach known as Autoregressive Integrated Moving Average (ARIMA) model to predict yield of given dataset. The research uses data analysis tools and packages to attain and compare the results from the proposed methods using R programming language. These data tools help to analyse the various ARIMA models and generates comparative analysis. The study uses different performance metrics and graphs to analyse the result. The detailed and comparative analysis is performed between observed values and predicted values for yield in tonnes per hectare.

4. CROP YIELD PREDICTION SYSTEM – DESIGN AND ARCHITECTURE

The data module, data pre-processing module, data analysis module, data processing (yield prediction algorithm) module, and user interface module make up the architecture of the crop yield prediction system. This data module offers the necessary database for forecasting price for the following twelve months [5]. The database contains various types of data that require further processing through data cleaning and validation of data types before being made available for subsequent analysis. Then, using a suitable analytical technique, predict the commodity's yield based on the available data. The system's ultimate result will be displayed in a user interface that is convenient for the end user. The Fig.1 depicts the crop prediction system's step-by-step method [6].

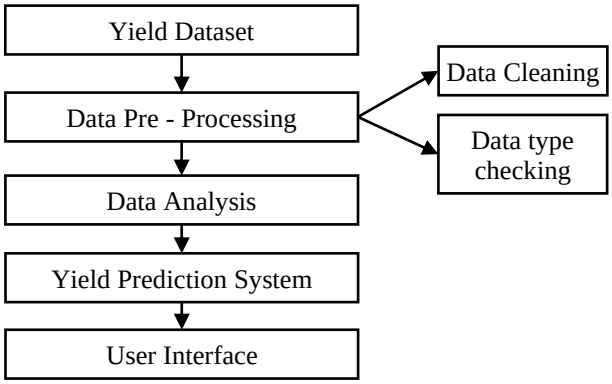

Fig.1. Architecture of Crop Yield Prediction System

5. FLOW CHART DEFINING THE YIELD PREDICTION SYSTEM

The dataset has been sourced from the Indian official website's Open Government Data (OGD) platform [4]. The data is downloaded from the year 1997 to 2021, Puttur Taluk of Dakshin Kannada district. The dataset has four attributes, all four attributes are numerical attributes. It covers information such as year, Area in terms of Hectares, Production in terms of tonnes, and yield in the form of tonnes per hectare. This dataset provides year-wise yield data for areca nuts, measured in tonnes per hectare, specifically for Puttur Taluk within the D.K district. During the data pre-processing stage, the dataset contains missing values for many years. These missing values are addressed by employing the linear regression method for imputation [7]. Next step is need to check for stationary, it is a hypothesis test in which the null hypothesis is that the sequences is non-stationary, and the alternate hypothesis is set to the sequence is stationary. If time series dataset is not stationary, then need to difference it to convert it into stationary series [8]. For the yearly areca nut yield series, we have plotted Line plots, autocorrelation function (ACF), and partial autocorrelation function (PACF) correlograms to finding out the stationarity of the yield data over time interval. In next step Based on the ACF and PACF after differencing the original series, we may decide the order of our ARIMA model [9]. Once the prototype is developed, it is required to develop other models to observe relevant coefficients, fluctuations, logarithmic probability density, Akaike's information criterion (AIC), and Bayesian (BIC) information criterion statistics [10]. Tabulated estimated statistic values are utilized for determining the most suitable model. During implementation stage, based on selected model apply ARIMA approach to find out predicted yield for next 4 years. The predicted result of the next 3 years is obtained in tabular form and exposed in the form of a graph for clear visibility.

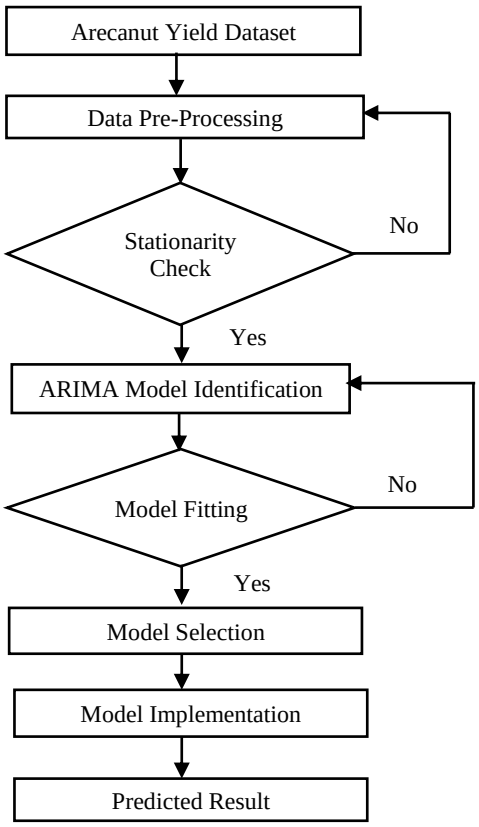

Fig.2. Flowchart for yield prediction using ARIMA model

6. STATIONARITY CHECK IN TIME SERIES

The stationary series is one whose statistical components like mean, mode, median, variance will not vary with time. The series

excluded with trend and seasonal components is stationary series. In time series data need to be tested for stationarity to achieve the desired result after applying any approach to process the data. In ARIMA model need to determine the number of differences to perform to make stationary series, then only it will forecast the predicted values. We cannot forecast non stationary data series. So, it is very much required to perform stationarity check especially in time series dataset which is preferred by any statistical method to predict future values. In statistics we use two main approaches for checking stationarity. The very first approach is Augmented Dickey-Fuller test (ADF test) and Kwiatkowski-Phillips-Schmidt-Shin test (KPSS). The very frequently using method is to plot the data series and observing the pattern for trend and seasonal components.

ADF test: This approach performs fast observation on given data and provides evidential result for stationarity in the data series. It is also called as unit root test which is performed based on the hypothesis. Where null hypothesis H_0 denotes the series is non-stationary and has unit root. Whereas alternative hypothesis H_A denotes the series is stationary and no unit root exist. The condition to reject the null hypothesis is that p value should be less than 0.05 indicates series is stationary series. The DFT considers auto regressive model and optimizes the information criteria with many lag values. It examines the null hypothesis with α =1is coefficient of first lag on y.

- Null hypothesis H0: alpha =1
- $y(t\text{-}1)$ = lag 1 of time series.
- £$Y(t\text{-}1)$ = first difference of the series at t-1.
- The $Y(t\text{-}1) - 1$ implies the existence of unit root means series is stationary.

In the same way ADT test is augmented form of DFT, where it expands the DFT equation by including higher order regressive model.

$$Y_t = x + \beta_t + a y_{t-1} + \acute{O}_1 \delta y_{t-1} + \acute{O}2 \delta y_{t-2} + \text{---} + \acute{O}_p \delta y_{t-p} + e_t$$

Since when the unit root exists and alpha =1, the value of p is less than 0.05, than it is required to reject the null hypothesis and considered as series is stationary.

KPSS test: It is also the approach for checking stationarity in series over deterministic way it is interchangeable with ADF test. Here the null hypothesis is considered as the series is stationery and interpretation of p value is opposite of ADF test. Here the p value is less than the significant level of 0.05, then series is said to be non-stationary. So it says that,

- Null hypothesis H_0: has no unit root and stationery
- Alternative hypothesis H_A: has unit root and not stationary series.

Consider the test fails to reject null hypothesis than it has to provide proper evidence to prove it has no unit root and trend stationary.

i.e. test_statistic < critical_value<0.05 then fail to reject H_0

if test_statistic > critical_value<0.05 then reject H_0

Here both ADF test and KPSS test are used for checking stationarity and has ambiguity to choose for implementation. It is suggested to use both the test and based on the result decide the series is truly stationary. There are four different use cases:

- The result of both test is stationary, then the series is stationary.
- The result of both test is non-stationary, then the series is non stationary.
- If ADF test result yields non stationary and KPSS test result yields stationary, then the series has trend stationary. The trend component must be removed from stationary series.
- If ADF test result yields stationary and KPSS test result yields non stationary than series is difference stationary. The differencing has to be performed to make it stationary.

7. IMPLEMENTATION

The yield data series of Puttur taluk contains year wise yield set from the year 1997 to year 2022. For the model selection, in the beginning we must check stationarity of the yield data series [11]. We need to draw the time plot of the yield data series which exhibits the trend of the series with different time interval. The Fig.3 shows the year wise time plot of the yield data series. If the time plot of the yield series shows huge declining pattern or surge pattern, then series is not a stationary series.

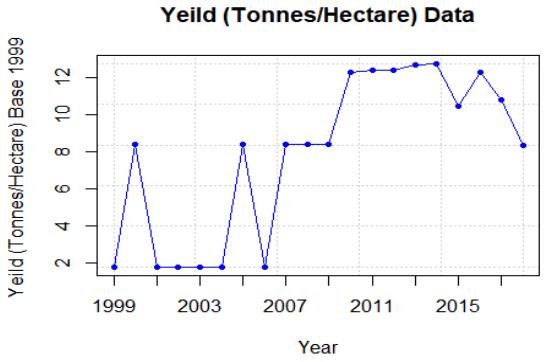

Fig.3. Time plot for yield data series

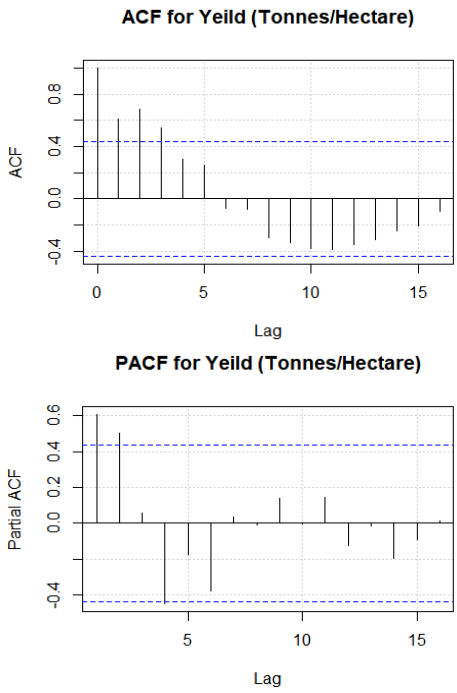

Fig.4. ACF and PACF of yield data series

The data plot of yield data series indicates a clear rising and declining trend from 1997 to 2022. So here series has a trend component implies it is not a fixed series. Additionally, it is recommended to utilize the ACF and PACF of the original dataset to assess stationarity [12]. The Fig.4 below illustrates the correlation functions of the yield data series.

In Fig.4 it is clearly indicates that ACF is gradually decreasing as number lag increases and one substantial spike that exceeds the standard error (SE) band shows the yield data set is not fixed. It is clear sign that need to perform first differencing to convert original series in to stationary series [13]. The Fig.5 illustrates the corelation functions of the yield data after performing the differencing.

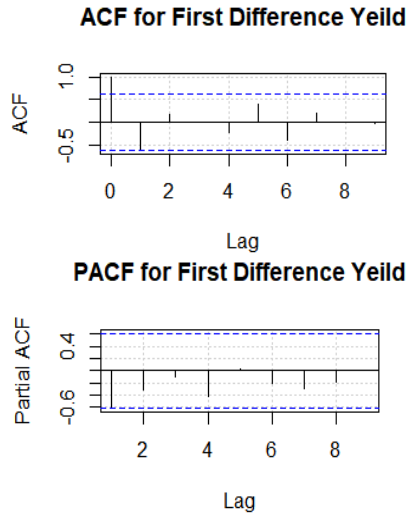

Fig.5. ACF and PACF of yield data series after differencing

Based on the corelation functions after differencing the original series helps to decide order of ARIMA model. Based on the ACF of first differencing the proportion of falloff is much faster which changes from positive autocorrelation at lag 5. In this case d=1, since we perform the first difference to transform original series in to stationary series [14]. In Fig.5 the ACF shows declining pattern in higher lags and significant spike at lag 1. So based on PACF of Fig.5 the order of p=1 and based on declining pattern of ACF order of q=0. To obtain more conclusive evidence and accuracy, several alternatives such as ARIMA (0,0,0), ARIMA (1,1,0), ARIMA (1,1,1), ARIMA (1,0,1), ARIMA (0,0,1), ARIMA (1,0,0), ARIMA (2,2,2), ARIMA (0,2,2), ARIMA (2,1,2), are considered.

To obtain more conclusive evidence and accuracy of the model, several alternatives of the approach is considered [15]. We compared different alternative models based on the lowest AIC value, highest log-likelihood, largest significant coefficient, and lowest RMSE value. The results are presented in Table.2.

Table.2. Comparison of alternative models with estimation criteria for yield series data

Alternative models/ Estimation Criteria	AIC	Log-likelihood	RMSE
ARIMA (0, 0, 0)	118.83	-57.41	4.270611
ARIMA (1, 1, 0)	98.98	-47.49	2.834469
ARIMA (1, 1, 1)	100.78	-47.39	2.819735
ARIMA (1, 0, 1)	109.34	-50.67	2.98929
ARIMA (0, 0, 1)	115.94	-54.97	3.761772
ARIMA (1, 0, 0)	110.75	-52.37	3.275552

In the Table.2, it is observed that ARIMA (1,1,0) model beats all other models with lowest AIC value, highest log-likelihood value and lowest RMSE value [16]. So, all the criteria are in the favor of ARIMA (1,1,0). So, we take ARIMA (1,1,0) for the diagnostic check. In addition, based on Ljung-Box test, we accepted the null hypothesis indicated that the residuals are white noise [17]. Thus, it can be identified that the suitable model based on ACF and PACF is ARIMA (1, 1, 0).

8. RESULT ANALYSIS

Here the year wise predicted yield of the arecanut is calculated from the year 1997-98 to 2025-26. The first year's dataset is considered as training dataset and actual prediction is done after the year 1998-99 [18]. The Table.2 represents the actual yield and predicted yield in tonnes per hectare of arecanut and forecast is done based on ARIMA (1, 1, 0) model for next 4 years 2022 to 2026.

Table.3. Prediction of arecanut yield

Year	Actual Yield (Tons/Hectare)	Predicted Yield (tons/Hectare)
1997-98	1.78	
1998-99	8.37	6.394502
2000-01	1.78	3.167138
2001-02	1.78	2.198937
2003-04	1.78	1.908478
2004-05	1.78	1.821321
2005-06	8.37	6.405702
2006-07	1.78	3.170508
2007-08	8.37	6.810462
2008-09	8.37	7.902449
2009-10	8.37	8.230048
2010-11	12.29	11.06914
2011-12	12.37	11.97632
2012-13	12.37	12.24847
2013-14	12.67	12.54494
2014-15	12.73	12.67547
2015-16	10.45	11.11726
2016-17	12.26	11.91452
2017-18	10.80	11.13498
2018-2019	8.32	9.8677
2019-2020	8.21	9.8677
2020-2021	9.35	8.903
2021-2022	9.21	9.5047
2022-2023		9.1294

2023-2024		9.3635
2024-2025		9.2175
2025-2026		9.3085

The Table.3 shows the predicted arecanut yield in terms of tonnes per hectare of puttur taluk from the year 1997-98 to 2025-26. The predicted yield of the year 2022, 2023, 2024, 2025 and 2026 is represented in Table.3. The Table.3 depicts the comparison of actual and predicted results of yield of arecanut in tonnes per hectare. It depicts that from year 1997-98 to year 2021-22 the actual yield and predicted yield shows similar value for ARIMA (1, 1, 0) model. The predicted yield for the year 2022-23, 2023-24, 2024-25 and 2025-26 indicates in future the production of arecanut will decreases in small amount.

9. CONCLUSION

It clearly shows here the ARIMA model is very suitable model for time series dataset to find the short-term prediction. In this experiment ARIMA (1,1,0) model is best model which gives accurate result and meets all performance criteria. According to this experiment the future production of arecanut for next 4 years shows gradually decreasing pattern, which is predicted around 9 to 9.5 tonnes per hectare. So this type of prediction helps all arecanut stakeholders to take appropriate decision based on chance of occurring future fluctuation in production. The result achieved in this experiment is holds good only in normal situations and does not applicable on abnormal condition.

REFERENCES

[1] D.D. Bhavani and R.B.S. Bharati, "An Efficient Method to Incorporate Precision Farming in Indian Agriculture Using Robotics and Internet of Things", *International Journal of Research in IT and Management*, Vol. 6, No. 9, pp. 1-12, 2017.

[2] S. Rajeswari and K. Rajakumar, "A Smart Agricultural Model by Integrating IoT, Mobile and Cloud-Based Big Data Analytics", *Proceedings of International Conference on Intelligent Computing and Control*, pp. 1-5, 2017.

[3] H. Kumar and T. Menakadevi, "A Review on Big Data Analytics in the Field of Agriculture", *International Journal of Latest Transactions in Engineering and Science*, Vol. 1, No. 4, pp. 1-10, 2018.

[4] 11 About Variety Wise Market Price, Available at https://data.gov.in/resources/variety-wise-daily-market-prices-arecanutbetelnutsupari-2020, Accessed in 2021.

[5] 11 About Agricultural Commodity Price, Available at http://agriexchange.apeda.gov.in/india%20production/India_Productions.aspx?hscode=1092, Accessed in 2021.

[6] H.C. Co and R. Boosarawongse, "Forecasting Thailand's Rice Export: Statistical Techniques vs. Artificial Neural Networks", *Computers and Industrial Engineering*, Vol. 53, No. 4, pp. 610-627, 2007.

[7] G.P.R. Rao and D.M. Srinivas, "Large Scale Farming Analysis with the Help of IOT and Data Analytics", *International Journal of Advanced Multidisciplinary Scientific Research*, Vol. 2, No. 3, pp. 1-13, 2019.

[8] M. Kumar and M. Nagar, "Big Data Analytics in Agriculture and Distribution Channel", *Proceedings of International Conference on Computing Methodologies and Communication*, pp. 384-387, 2017.

[9] About ARIMA Time Forecasting Model, Available at https://otexts.com/fpp2/arima.html, Accessed in 2023.

[10] About Mathematical Theory model of ARIMA, Available at https://people.duke.edu/~rnau/411arim.html, Accessed in 2023.

[11] About Time Series Data Analytics, Available at https://www.analyticsvidhya.com/blog/2021/11/performing-time-series-analysis-using-arima-model-in-r/, Accessed in 2023.

[12] Z. Chen, H.S. Goh and X.Y. Liew, "Automated Agriculture Commodity Price Prediction System with Machine Learning Techniques", *Proceedings of International Conference on Machine and Deep Learning*, pp. 1-12, 2021.

[13] Y.N. Havaldar and A. Kamei, "Forecasting of Areca Nut (Areca catechu) Yield using Arima Model for Uttara Kannada District of Karnataka", *International Journal of Science and Research*, Vol. 3, No. 7, pp. 1002-1009, 2014.

[14] L. Narsimhaiah, K. Sinha and P. Pandit, "Modeling and Forecasting of Areca Nut Production in India-Vision", *International Journal of Current Microbiology and Applied Sciences*, Vol. 8, No. 11, pp. 728-738, 2020.

[15] S.A. Mulla and S.A. Quadri, "Crop-Yield and Price Forecasting using Machine Learning", *International Journal of Analytical and Experimental Modal Analysis*, Vol. 12, pp. 1731-1737, 2020.

[16] P. Samuel, D. Ramanika and N.A. Kumar, "Crop Price Prediction System using Machine Learning Algorithms", *Journal of Software Engineering and Simulation*, Vol. 6, No. 1, pp. 14-20, 2020.

[17] P.S. Rachana, N. Shruthi and R.S. Kousar, "Crop Price Forecasting System using Supervised Machine Learning Algorithms", *International Research Journal of Engineering and Technology*, Vol. 6, pp. 4805-4807, 2019.

[18] X. Pham and M. Stack, "How Data Analytics is Transforming Agriculture", *Business Horizons*, Vol. 61, No. 1, pp. 125-133, 2018.