**Time series analysis of Agricultural data: WHEAT, RICE AND MAIZE**

**Aditi Singh1, Tinni Chaudhuri2**

**Pankaj Kumar3**

1 Research Scholar, Department of Statistics, Amity University, Kolkata

2 Assistant Professor, Department of Statistics, Amity University, Kolkata

3 Assistant Professor, Department of Statistics, Amity University, Kolkata

**1.1 INTRODUCTION**

Wheat (Triticum Aestivum), the world's most widely grown cereal crop, is a member of the Graminae (Poaceae) family of Triticum. Because of its vast acreage, excellent production, and significant position in the international food grain trade, it has been dubbed the "King of Cereals." Wheat is used to making bread, chapatti, porridge, flour, suji, and other foods. The word "wheat" comes from a variety of places, including the English, German, and Welsh languages. Wheat has a strong nutritional profile, with 12.1 percent protein, 1.8 percent lipids, 1.8 percent ash, 2.0 percent reducing sugars, 6.7 percent pentosans, 59.2% starch, and 70% total carbohydrates, and delivers 314 calories per 100 grams. It also contains a lot of minerals and vitamins.

Maize (Zea mays L.), a cereal grain that also belongs to the grass family Poaceae and the genus Zea, is widely grown. Corn is another name for it. In many places of the world, maize has become a staple diet. Maize is used for corn ethanol, animal feed, and other maize products, such as corn starch and corn syrup, in addition to being consumed directly by humans (typically in the form of masa). The English term "maize" comes from the Spanish word maize. Maize's nutritional composition includes 96 calories, 73 percent water, 3.4 grams of protein, 21 grams of carbs, 4.5 grams of sugar, 2.4 grams of fiber, and 1.5 grams of fat. Corn may be high in a variety of vitamins and minerals. Notably, the amount is highly variable depending on the corn type.

The scientific name for rice is Oryza. The most frequent species is Oryza sativa. Rice is a staple food in over 100 nations throughout the world. Rice is served with more than one meal every day in some households. This high-calorie starchy grain is often inexpensive, making it accessible to all and a necessary component of many diets. Rice has the following nutritional profile: 130 calories, 28.7 grams of carbohydrate, 2.36 grams of protein, and 0.19 grams of fat.

Crop yield forecasting is crucial for proper planning and policymaking in the agriculture sector of the country. The present study was undertaken to investigate the feasibility of estimating the productivity of rice, maize, and wheat crops based on production variables using past yield records of India.

**1.2 METODOLOGY AND DATASET**

Maize and wheat crop yield data for the period of the recent 68 years (1950-51 to 2017-18) and Rice crop yield data for the period of 67 years (1951-52 to 2017-18) were used to develop yield forecasting models. The data was collected from the yearend book *Agricultural Statistics at a Glance 2018* released by the *Government of India, Ministry of Agriculture & Farmers Welfare Department of Agriculture, Cooperation & Farmers Welfare Directorate of Economics and Statistics.* The variable year contains the financial year whereas the production is in terms of million tonnes. Python software was used for the analysis of data.

A **time series** (or dynamic series) is a series of data that are ranked by the time of occurrence and have the same statistical indicators of value. The primary goal of time series analysis is to forecast the future using previous data. In practice, numerous factors influence the change of the time series, some of which are long-term and definitive, resulting in some trends and regularities in the time series, while others are short-term and indecisive, resulting in some irregular changes in the time series. Changes in time series can be loosely divided into three categories:

1. Trend change refers to a phenomenon's tendency to display a steady increase, steady decline, or steady trend in a particular direction throughout time.
2. Seasonal changes (cycle changes) refer to the phenomena of seasonal effect, which is based on a set cycle that shows cyclical swings.
3. The phenomenon of irregular factors in the impact of irregular fluctuations is referred to as stochastic changes.

In most cases, time series is a result of a mix of the elements listed above.

The term "time series forecasting" refers to a form of forecasting that uses the variable's past data. It is the direct application of the continuity principle by determining the historical pattern of variables and assuming that this pattern will continue to hold in the future. Short, medium and long-term forecasting can all be done using the time series forecasting method.

**Autoregressive Integrated Moving Average (ARIMA)** was fitted after determining the stationarity of the differenced data with the ADF test. An autoregressive integrated moving average, or ARIMA, is a statistical analysis model that uses time-series data to either better understand the data set or to predict future trends. A statistical model is autoregressive if it predicts future values based on past values.

Each of the components of an ARIMA model can be understood by outlining them as follows:

* Autoregression (AR): refers to a model whose observed values of a variable regresses on the lagged, or previous observed values of the same variable.
* Integrated (I): denotes the differencing of raw observations in order to make the time series stationary (i.e., observations are replaced by the difference between the present observation and the previous observation).
* Moving average (MA): refers to the dependency between an observation and a residual error when a moving average model applied to lagged observations.

ACF is a (complete) auto-correlation function that returns auto-correlation values for any series with lagged values. These values are plotted beside the confidence band. It expresses how closely the series' current value is related to its previous values. Components of a time series include trend, seasonality, cyclic, and residual. Because ACF incorporates all of these factors while determining correlations, it is known as a 'complete auto-correlation plot.' PACF stands for partial autocorrelation function. Essentially, instead of finding present with lags correlations like ACF, it finds a correlation of the residuals (which remain after removing the effects that are already explained by the earlier lag(s)) with the next lag value, thus 'partial' rather than 'complete' because we remove previously found variations before finding the next correlation.

The PACF plot is used to discover the optimum features or order of the AR process since it removes differences explained by prior lags, leaving only the relevant features.We use the ACF plot to discover the optimum features or order of the MA process. Because the MA process has no seasonal or trend components, we only receive the residual relationship with the lags of time series in the ACF plot. ACF acts as a partial plot.

**1.3 RESULT AND DISCUSSION:**

**WHEAT**

The initial data plotted, shows a trend but no seasonality which was expected as our data is yearly.



*Figure 1: Wheat Yearly Production data*



Figure 2: Decomposition of wheat data

After splitting the data into train and test data, train data was then decomposed using the multiplicative model.

As we can see there is no seasonality in the data. After differencing the data with lag order 1, the ADF test p-value came out to be -1.202658e+01 which is less than 0.05. Thus, we reject the null hypothesis and accept the alternative hypothesis that the differenced test data with lag order 1 is stationary.

We now plot the ACF and PACF graphs for the stationary data.



Figure 4: PACF Graph

*Figure 3: ACF Graph*

According to the ACF and PACF graph, we will now test ARIMA fitting with q value 1 and p value ranging from 1 to 10. We select the model with the least AIC and BIC values along with keeping in check whether the coefficients are significant or not, which in this case is obtained for the model with order (2,1,1).



*Figure 5: ARIMA (2,1,1) Summary*

The density plot for the residual of the model is as follows:



Figure 6: Density graph for residual

Thus, the residual follows a normal distribution with (0, 𝛔2). After fitting the model on test data the root mean square error (RMSE) value for the same is 6.050464 which is not very informative by itself, but you can use it to compare the fits of different ARIMA models.

Model fit on test data:



Figure 7: Model fit graph

The production forecast for 2018-19 to 2022-23 was obtained to be 92.27415195, 93.45132593, 94.65766924, 95.86542423, 97.08562877, 98.31273423, 99.54835726 (Metric Tonnes) respectively.

**RICE**

The initial data plotted, shows a trend but no seasonality which was expected as our data is yearly.



*Figure 8: Rice Yearly Production Data*

After splitting the data in train and test data, train data was then decomposed using multiplicative model.

Figure 9: Decomposed data

As we can see there is no seasonality in the data. After differencing the data with lag order 1, the ADF test p-value came out to be 5.833169e-08 which is less than 0.05. Thus, we reject the null hypothesis and accept the alternative hypothesis that the differenced test data with lag order 1 is stationary.

We now plot the ACF and PACF graphs for the stationary.

Figure 11: PACF graph

*Figure 10: ACF graph*

According to the ACF and PACF graph, we will now test ARIMA fitting with q value 1 and p value ranging from 1 to 16. We select the model with the least AIC and BIC values along with keeping in check whether the coefficients are significant or not, which in this case is obtained for the model with order (1,1,1).



*Figure 12: ARIMA(1,1,1) Summary*

The density plot for the residual of the model is as follows:



Thus, the residual follows a normal distribution with (0, 𝛔2). After fitting the model on test data the root mean square error (RMSE) value for the same is 4.050464 which is not very informative by itself, but you can use it to compare the fits of different ARIMA models.

Model fit on test data:

Figure 13: Density graph of the residual



Figure 14: Model fit graph

The production forecast for 2017-18 to 2021-22 was obtained to be 110.00111421, 111.29413312, 112.5871547, 113.88017586, 115.17319709 (Metric Tonnes) respectively.

**MAIZE**

The initial data plotted, shows a trend but no seasonality which was expected as our data is yearly.



*Figure 15: Maize yearly production data*

After splitting the data into train and test data, train data was then decomposed using the multiplicative model.



Figure 16: Decomposed data

As we can see there is no seasonality in the data. After differencing the data with lag order 1, the ADF test p-value came out to be 2.437900e-07 which is less than 0.05. Thus, we reject the null hypothesis and accept the alternative hypothesis that the differenced test data with lag order 1 is stationary.

We now plot the ACF and PACF graphs for the stationary data.



Figure 18: Decomposed data

8

*Figure 17: ACF Graph*

According to the ACF and PACF graph, we will now test ARIMA fitting with q value 1 and p value ranging from 1 to 3. We select the model with the least AIC and BIC values along with keeping in check whether the coefficients are significant or not, which in this case is obtained for the model with order (2,1,1).



*Figure 19: ARIMA (2,2,1) Summary*

 The density plot for the residual of the model is as follows:



Figure 20: Density graph of the residual

Thus, the residual follows a normal distribution with (0, 𝛔2). After fitting the model on test data, the root means square error (RMSE) value for the same is 1.195115 which is not very informative by itself, but you can use it to compare the fits of different ARIMA models.

Model fit on test data:



Figure 21: Model fit graph

The production forecast for 2018-19 to 2022-23 was obtained to be 27.98793671, 28.90664384, 30.01304574, 31.05604598, and 32.0560779 (Metric Tonnes) respectively.

**1.4 CONCLUSION**

The paper analyses the trend in terms of the production of wheat, maize, and rice across India. The growth has been examined by production from 1950-51 to 2017-18 for wheat and maize, and 1951-52 to 2017-18 for rice. The purpose of this paper is to fit an ARIMA model on the production data of wheat, rice, and maize, and predict the future forecast for the next five years. Results indicate that the fitted ARIMA model predicts the data with a root mean square error of 6.050464, 4.050464 and 1.195115 for wheat, rice and maize respectively.

* The production forecast of wheat for 2018-19 to 2022-23 was obtained to be 92.27415195, 93.45132593, 94.65766924, 95.86542423, 97.08562877, 98.31273423, 99.54835726 (Metric Tonnes) respectively.
* The production forecast of rice for 2017-18 to 2021-22 was obtained to be 110.00111421, 111.29413312, 112.5871547, 113.88017586, 115.17319709 (Metric Tonnes) respectively.
* The production forecast of maize for 2018-19 to 2022-23 was obtained to be 27.98793671, 28.90664384, 30.01304574, 31.05604598, and 32.0560779 (Metric Tonnes) respectively.

**1.5 REFERENCE**

Online References:

* [*https://towardsdatascience.com/significance-of-acf-and-pacf-plots-in-time-series-analysis-2fa11a5d10a8*](https://towardsdatascience.com/significance-of-acf-and-pacf-plots-in-time-series-analysis-2fa11a5d10a8)

Papers Referred:

1. *Yield forecasting of rice and wheat crops for eastern Uttar Pradesh* by R.S. SINGH, CHANDRABHAN PATEL, M.K. YADAV and K.K. SINGH1, Department of Geophysics, B.H.U., Varanasi (UP) 221 005, India Meteorological Department, Mausam Bhavan, New Delhi

E-mail: rsingh\_61@yahoo.co.in

1. *A Statistical Study of Trends of Wheat Production in Districts of Eastern Uttar Pradesh*, India Anvesha Singh, Sri Sunil Kumar, Gopal Krishna Tiwari, Laxmi Devi and Babulal Department of Agricultural Statistics, Acharya Narendra Dev University of Agriculture and Technology, Kumarganj – Ayodhya (UP), 224-229, India