

Wheat Production Forecasting in Bangladesh Using Deep Learning Techniques

BY

KOHINOOR HAQUE

ID: 213-25-066

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Computer Science and Engineering

Supervised By

Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

SEPTEMBER 2022

APPROVAL

This Thesis titled "Wheat Production Forecasting in Bangladesh Using Deep Learning Techniques" submitted by Kohinoor Haque, ID No: 213-25-066 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 19-01-2021.

BOARD OF EXAMINERS



Chairman

Dr. Sheak Rashed Haider Noori, PhD
Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



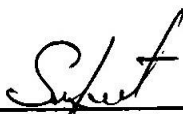
Internal Examiner

Dr. Moushumi Zaman Bonny
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Internal Examiner

Md. Sazzadur Ahamed
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



External Examiner

Md. Safaef Hossain
Associate Professor & Head
Department of Computer Science and Engineering
City University

DECLARATION

We hereby declare that, this thesis has been done by Kohinoor Haque under the supervision of, Abdus Sattar, Assistant Professor, Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for the award of any degree or diploma.

Supervised by:



Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:



Kohinoor Haque
ID: 213-25-066
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

From the very beginning, I would want to express my unwavering gratitude to Allah-like for bestowing upon me the ideal blessing that has enabled me to reach the point of completing my master's thesis. I really grateful and wish my profound our indebtedness to **Mr. Abdus Sattar, Professor and head Department of CSE, Daffodil International University**. I would like to express my gratitude to my manager for providing me with the necessary direction to do this outstanding piece of research work for the evaluation test that made use of facts regarding person-to-person communication. I would not have been able to complete this investigation endeavor to the letter without his assistance and direction. He made all of the associated assets and crucial data that we needed available to us so that we could complete this investigation for sentiment analysis. I'd also like to express my gratitude to my co-director for helping us see this project through to completion.

I would like to express my heartiest gratitude to Prof. Mr. Abdus Sattar and Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

Wheat is an important crop for Bangladesh, which is both a producer and a consumer of grain. Bangladesh's agricultural sector is vital to the country's economy and employment levels. Wheat is the most important winter crop that is grown during the winter (Rabi) season. Wheat is also an essential food staple. Wheat yield and production statistics in Bangladesh are typically released by the local crop reporting administration several months after harvest has taken place. The field data were gathered by hand from a predetermined list of villages to provide the basis for the statistical estimates. Forecasts of early season wheat production enable improved planning for wheat transactions on the global market, the maintenance of adequate stocks, the informing of policymaking, the setting of support prices, and an increase in market efficiency. The long-term viability of Bangladesh's wheat industry as well as its susceptibility to market swings is the topic that will be investigated in this study. According to the results of the experiment, Bangladesh's wheat crop area, production, and yield are all on the rise. The experiment was conducted in Bangladesh. In order to develop the model and obtain an estimate of the forecasting behavior, the ARIMA model methodology was utilized.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
List of Figure	v
List of Tables	v
CHAPTER	
CHAPTER 1: INTRODUCTION	1-4
1.1 Introduction	1
1.2 Motivation	1
1.3 Rational of the study	2
1.4 Research Data	3
1.5 Expected Outcome	3
1.6 Report Layout	3-4
CHAPTER 2: BACKGROUND STUDIES	5-8
2.1 Introduction	5
2.2 Related Works	5-8
2.3 Research summary	8
CHAPTER 3: RESEARCH METHODOLOGY AND IMPLEMENTATION	9-18
3.1 Introduction	9
3.2 Implementation	9-11
3.3 Data collection	11-12
3.4 Importance of stationarity	12-13
3.5 Time Series Analysis	13-15
3.6 ARIMA	15-18

CHAPTER 4: EXPERIMENTAL RESULT AND DISCUSSION	19-23
4.1 Introduction	19
4.1 Back-end Design	19-20
4.3 Model Identification and Parameter Estimation	20-21
4.4 Model Performance and Evaluation Matric	22-23
4.5 Analysis of Forecasting Results	23
CHAPTER 5: CONCLUSION	24-25
5.1 Conclusion	24
5.2 Future Work	25
REFERENCE	26-27

LIST OF FIGURES

FIGURE	PAGE
Figure 1.2: Northern Region of Bangladesh	2
Figure 1.3: Annually Wheat Import Report	2
Figure 2.1: Year Wise Wheat Production of Bangladesh	5
Figure 3.2: Working Process with ARIMA	9
Figure 3.3: Data Entry in Spreadsheet	11
Figure 3.3.1: Visualization of Collected Data	12
Figure 3.4: Components of Time Series	12
Figure 3.4.2: Autocorrelation	14
Figure 3.4.3: Partial Autocorrelation	15
Figure 3.5: Working Process of ARIMA	16
Figure 4.3: Analysis Result	21
Figure 4.4: Original vs Predicted Result	22
Figure 4.4.1: Result of MSE	23

LIST OF TABLES

TABLES	PAGE
Table 1: Stationary Test Result	19
Table 2: Model Identification	20
Table 3: Results of Covariance	21
Table 4: Forecasting Results	23

CHAPTER 1

INTRODUCTION

1.1 Introduction

Bangladesh is a developing agricultural nation. The majority of her residents depend on agriculture, either directly or indirectly, for their living. 13.35% of the nation's Gross Domestic Product (GDP) is accounted for by agriculture (BBS,2020). Still, the country's economy's most significant sector remains agriculture, which continues to play a crucial role. By birth, Bangladesh has exceptionally fertile ground where a variety of crops may thrive readily. This nation produces a wide range of crops. One of the most significant cereal crops worldwide is wheat. In terms of cultivation and yield per hectare, it comes in #1. Most of the world's population lives on it, and around one-third of all people cultivate it. Since wheat is grown in Bangladesh during the Rabi season, it faces less competition from rice on available land. Along with other crops including vegetables, oil seeds, and pulses, wheat is one of the Rabi crops that are grown during the winter. The production of rice has been observed to be uncertain during the past few years due to a variety of natural catastrophes and production costs. The cultivation of wheat can be prioritized in this situation as a response to the food crisis. Due to a lack of irrigation infrastructure, farmers in the northwest of the country leave their land fallow during the Rabi season even though it could easily be planted in wheat.

1.2 Motivation

After rice, wheat is the second-most significant grain crop in Bangladesh. The importance of expanding the cultivation of this crop was acknowledged by the decision-makers of agricultural extension programs. Wheat has a great deal of potential for widespread cultivation as a high-value crop. Unfortunately, though, the rate of wheat cultivation has been declining over the past few years after initially trending upward. The majority of Bangladesh's wheat is grown in the northern Thakurgaon, Dinajpur, Rangpur, Panchagarh, and Lalmonirha districts. We must increase the rate at which wheat is grown in Bangladesh. Adapting new technologies, expanding farmer potential for wheat farming, using high-yield seeds, etc., can all change the cultivation process.



Figure 1.2: Northern Region of Bangladesh [1]

1.3 Rational of the study

The annual need for wheat is rising steadily in our nation as wheat-based foods like bread, bakery goods, and fast food items gain popularity. The amount of wheat imported by Bangladesh is likewise rising; it was 2039 MT in 2011 (BBAS, 2010-11), and after 10 years, it more than tripled to 6800 MT (BBS, 2020-21).

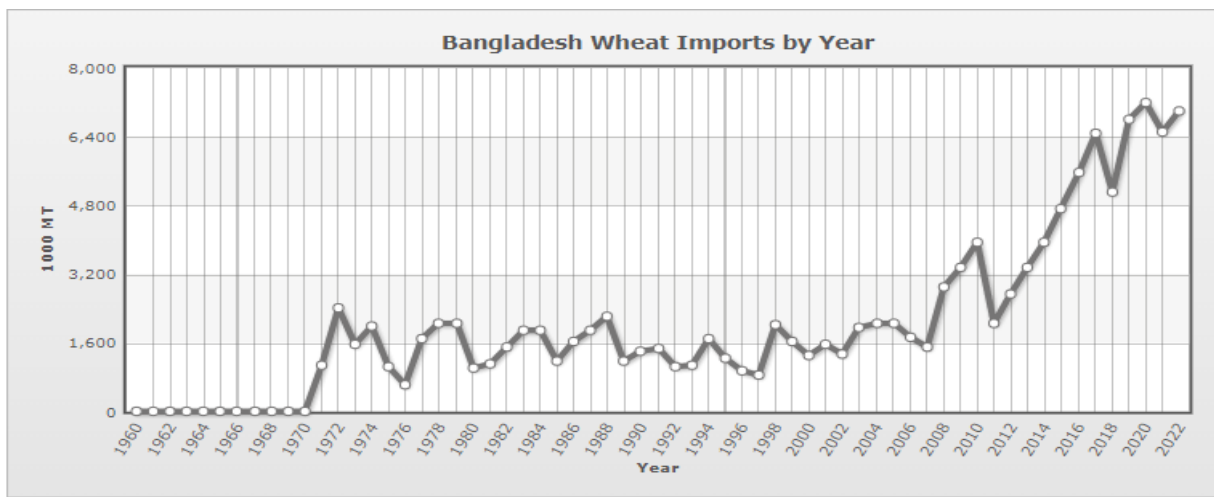


Figure 1.3: Annually Wheat Import Report [2]

Therefore, a sizable sum of local money must be spent on the importation of wheat. By accelerating the acceptance of wheat and the production of it as well, this might be reduced. Bangladesh must continue to make improvements. Therefore, the cultivation procedure needs to be improved.

1.4 Research Data

A coding sheet was used to code the data. By employing a computer system, the information gathered from the respondents was assembled, tabulated, and then analyzed in accordance with the study's goals. By using appropriate scoring systems, the qualitative data were transformed into quantitative form for analysis. We gather secondary information about the production of wheat in Bangladesh from the Bangladesh Wheat and Maize Research Institute.

1.5 Expected Output

Bangladesh relies heavily on wheat, making accurate predictions of wheat harvests crucial to the country's economy and food supply. Growing attention has been paid to the issue of developing a simple, quick, and accurate agricultural production forecast model at the administrative level through the integration of multi-source data and the application of machine learning techniques. Intensive manual inspections, remote sensing, or temperature data were applied in many earlier research, but their focus was on the entire crop growing cycle. However, it was unclear how picking a different time range would affect yield forecast. Using the key winter wheat producing areas of Bangladesh as an example, we segmented the entire growing period into four time periods and evaluated their respective prediction ability.

1. To assess wheat production through farming.
2. Fit a regression analysis to determine Bangladesh's wheat production.
3. Fit a time series model to guide decisions on Bangladesh's future wheat farming.
4. Make an effort to comprehend the production patterns using time series analysis.

1.6 Report Layout

First, the purpose for the research is discussed inside the preface (Chapter 1), which is what drove the author to tackle this particular issue statement in the first place. In this section, we will talk briefly about the goals of this study as well as an overview behind the work.

In the second part of the article, which is the existing literature portion (Chapter 2), we went over other works from the computer science field that dealt with the same kinds of problems. In addition

to that, the background investigation was conducted with the intention of determining the deficiencies that were present in earlier research.

The data collecting and conversion procedure that will be utilized in this project has been detailed in the data Methodology phase, which can be found in Chapter 3. This part also involves the extraction of features and comprehensive research on how each feature contributes to the overall conclusion. In Chapter 4 of the book, "Implementation and Model Selection," we explore the different forms of data, as well as which technique is superior for the various portions of patient data, and then we choose the appropriate algorithm.

In addition, the proposed models and a comparison analysis of the prediction rate among the various models were presented in the chapter titled "Results and Analysis." A presentation also serves to summarize the results of the study.

CHAPTER 2

BACKGROUND STUDIES

2.1 Introduction

The economy and employment of Bangladesh are greatly influenced by agriculture. In 2019, Bangladesh's economy's total Gross Domestic Product (GDP) was about 12.92% accounted for by agriculture, which also provided 50% of all jobs.

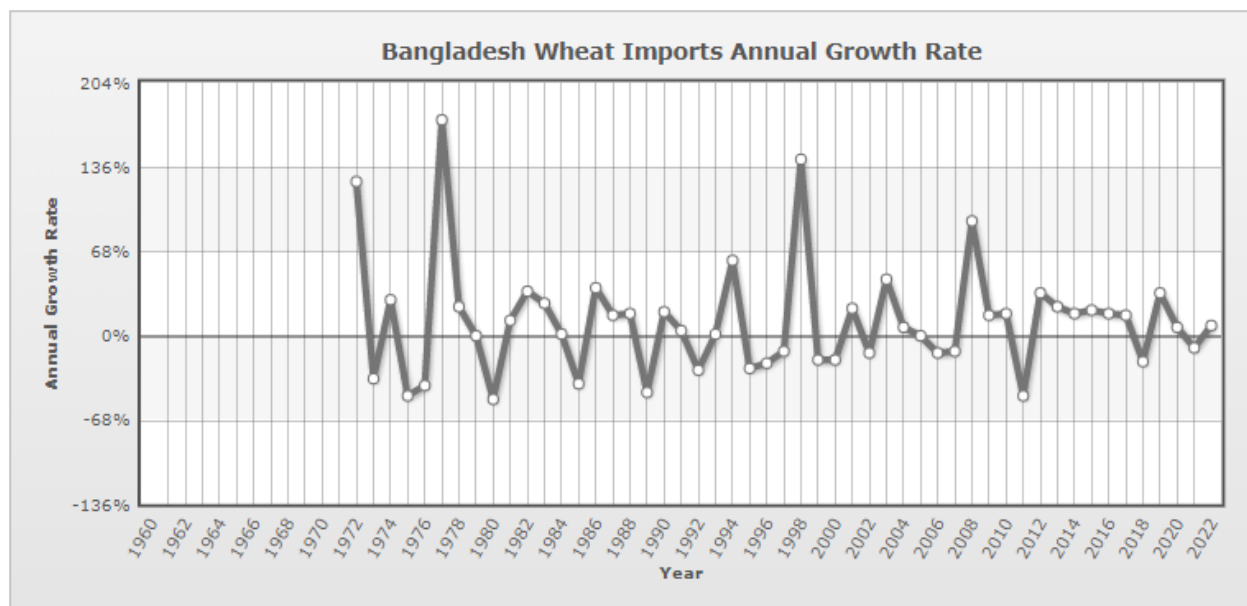


Figure 2.1: Bangladesh Wheat Imports Annual Growth Rate [2]

2.2 Related Work

The most significant cereal crop and the primary source of nutrition and energy in the daily diet of humans is wheat [4]. 16.6% of the world's wheat production is used to feed animals, compared to 66% for human consumption. In 2019, it produced 605.99 million tons worldwide on an area of 218 million hectares [5]. With a production of 1180 tons in 2020, Bangladesh was the major producer of wheat in close proximity to China. For about 33.3% of the population, wheat serves as both a primary food source and a major source of protein, niacin, and thiamine in the human diet [6]. Wheat (*Triticum* spp.) holds the top position in the world's food crop hierarchy [7]. One of the world's major producers and consumers of wheat is Bangladesh. It is only second to rice in

importance to Bangladesh agriculture. Products made from wheat flour, like chapatti, are essential for the staple diet in many parts of India, particularly in the north. Additionally, cattle are fed with wheat straw. The rate of population growth in the world, particularly in developing nations, is alarming. Care for this steadily growing human population continues to be a challenging task for national and international policymakers [8]. The potential population behaviour of the nations in an evolving situation should be known to the policymakers. They should also be aware of potential food and other product demand at the same time. Therefore, forecasting the production patterns of the major crops is crucial for policymakers to ensure the security of food and nutrition. The potential production scenarios for the major crops should be known to the policymakers [9].

Both agricultural development and instability continue to be hot topics in India and around the world. Even though there is a clear need to increase agricultural production or growth, there are a number of reasons why more production uncertainty is viewed negatively. Uncertainty affects farmers' income, their options for implementing high-wage technology, and their ability to invest in agriculture. And it raises the stakes for all food production. Low-income families are more vulnerable to market fluctuations due to instability in manufacturing, which also affects consumers and prices. Stable agriculture and food production is crucial to both effective food management and macroeconomic stability [10]. It is crucial to growing important crops using sustainable technology. After deciding what must be sustainable, it is possible to define sustainability clearly within the framework system. Agricultural systems function on multiple levels, from the local to the national to the international, making it difficult to set limits [11]. These include, but are not limited to, the soil-plant system, agricultural system, farming practices, agrarian environment, etc. In article [12], the researchers gathered a chronological dataset to conduct weather forecasting study. Nonlinear and conventional approaches both ineffective and inefficient because of the complex nature of atmospheric processes. As a result, coming back towards the prior chapter, deep learning appears to be the best match for finding and solving issues of this complexity. They utilize the following approach in their study. Preparation of the data: After experimenting with numerous strategies for dealing with missing data, the researchers opted to utilize the Spline Interpolation approach, which showed to be the most efficient of them.

Algorithm employed: The model addressed in this research examines performance differences due to different input styles. Meteorology information are available from a variety of local and

international organizations. These sources include aggregate data from many previous years, covering factors like maximum and lowest temp, prevailing winds, precipitation, wind speed, and so on. Working across all of these aspects, nevertheless, presents a major obstacle because of a lack of appropriate and powerful technology. For this reason, the experiment provided here could only test a single variable. To do this, we used artificial neural networks (ANN), one of the most powerful prediction-based research approaches. While most ANNs utilize an input-output ratio of 10:1, the approach shown here employs two unique ratios while maintaining the same dataset. The first test had a ratio of 4:1, whereas the second had a ratio of 19:1. Then, they were compared with one another. The supplied town's information had around 1,500 and 7,000 items, respectively. The input data was divided into three proportions: 8:1:1 for testing, training, and validation. Although it appears that combining data from several sites should result in quite accurate weather forecasts, particularly for larger regions, the actual experiment and testing demonstrated the opposite result. Upon combining the data, this ANN model's Mean Square Error rose considerably. This suggests that studies of the temperature in various regions and towns have very little in common with one another. The research also showed that the frequency of meteorological abnormalities and variations, however little, was higher in cities with relatively significant pollution levels. Employing non-traditional computerized methods, the model established in this study was a novel method for determining the status of climate change.

In their study [13], the authors demonstrate whether classifications like the Naive Bayes and Chi-Square algorithms could be employed for meteorological forecasting. The researchers developed a client-login-required web application with a functional graphical user interface (GUI). Audiences will input data (or parameters) such as the current weather forecast, humidity, temperature, wind speed, and direction, and so on. Next, it uses its ever-growing data could evaluate the input factors and make prognostications. The results of this application provide compelling evidence that data mining techniques could be suitable for use in meteorology.

A certain work presents a classification technique for climatology that illustrates the classification capabilities using Naive Bayes and Chi square techniques. This system has a well-designed graphical user interface and is a web-based application. To access the system, each user must input their login and password. Users will input data such as the current weather prediction, temperature, humidity, and wind speed. This application will utilize this variable to forecast the weather after

comparing the input data to the data records. As a consequence, classification (training) and prediction (testing), two key tasks, would be completed. Its outcomes demonstrated that these data mining techniques are suitable for forecasting weather.

Overall results of their own hypothesis as well as the materials they used are explained by the Chi Square test. A second test demonstrates that the results significantly deviate from what would be predicted from the training set's features. This model instructs the feasibility of the research in a context where a chi-square result with much more than 2 degree of variance is regarded meaningful. To understand the full picture for wheat production behavior for potential futures, the study of the total sustainable expansion of wheat production in the main wheat-growing states, including India, is necessary. Mexico's wind speed is predicted using the mixed ARIMA-ANN model [14]. A hybrid ARIMA was used in several studies [15, 16] for forecasting. In [17,18,19,20], predicting the wheat harvest using ARIMA and ARIMA-WNN hybrid models was discussed.

2.3 Research Summary

Forecasting and analyzing wheat production behavior is therefore required to address the issues of food demand and nutritional security. Although estimating models and predicting phenomena have a long history, their application, particularly in the field of agriculture, has become much more apparent over the past fifty years. The creation of the ARIMA methodology, followed by the availability and widespread use of computer software, gave it a boost. Later, various techniques entered the field of forecasting. Due to the complexity of time series, neither their linear nor nonlinear components should be studied in isolation because no time series can be considered to be entirely linear or nonlinear. The best methods for capturing the full behavior of these kinds of data series are hybrid forecasting techniques.

CHAPTER 3

RESEARCH METHODOLOGY AND IMPLEMENTATION

3.1 Introduction

To assess trends in data over time, time series plots are employed. Data from time series typically include elements like trend, seasonality, cycles, and outliers. Forecasting is a crucial tool for efficient and effective planning [21]. A key component of every scientific investigation is methodology. The research methodology should allow the researcher to get reliable data and appropriately analyze it in order to make informed judgments. This chapter's goal is to detail the methodologies and processes used to carry out the current investigation. The stages that are necessary for a project to be successful are listed below.

3.2 Implementation

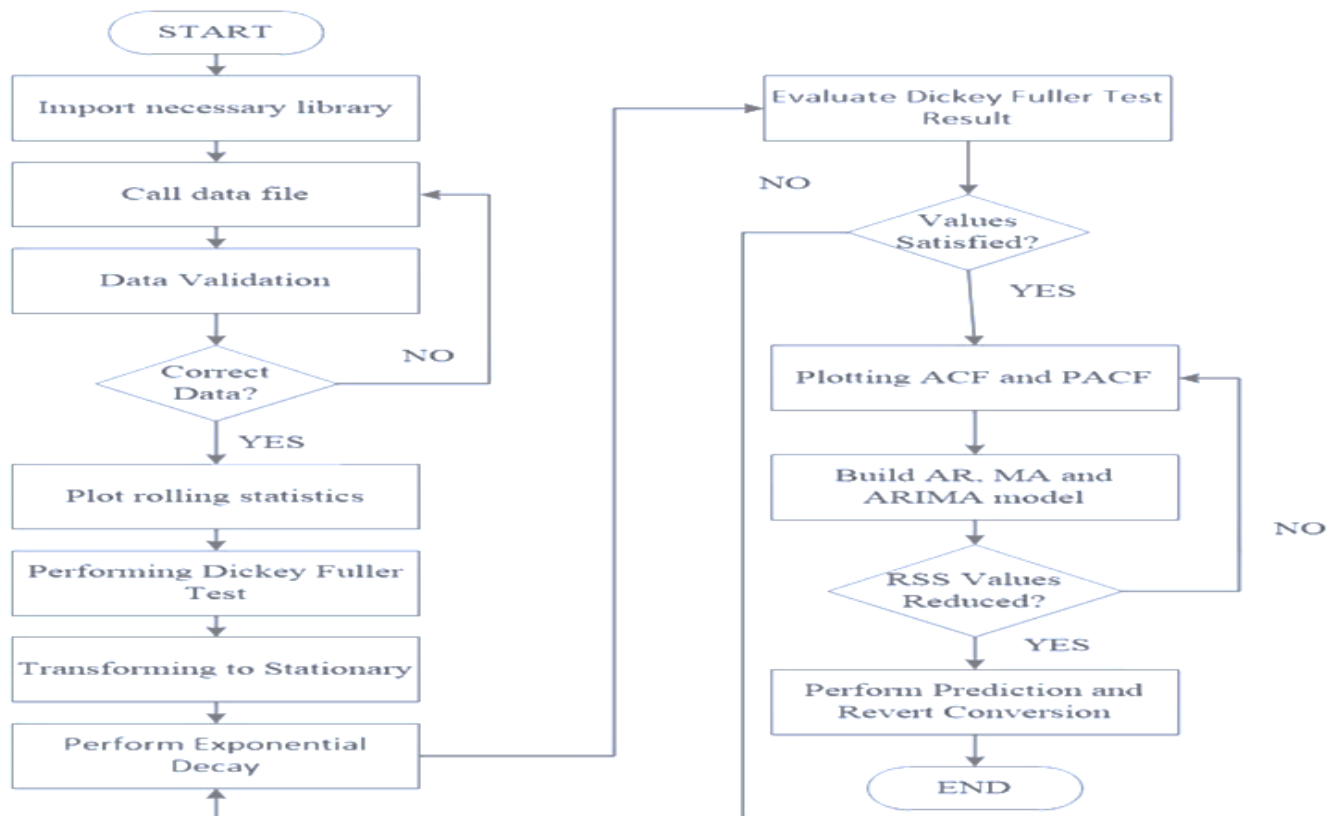


Figure 3.2: Working Process with ARIMA

A forecast is carried out through a series of steps. The Python library must be improved to meet the needs of data analytics. Google Colab was used to deploy the packages. Numpy, pandas, matplotlib, and rcParams are the packages.

Following the flow chart presented in Figure 2, the required Python packages were imported into the library. The following stage to accomplish is to retrieve the data source. The file contains a number of rejected applications dating back to the beginning of January 1949 and continuing through December 2018. Python commands used to check the legitimacy of the imported file. Python's head | tail command had been used to determine whether or not the import data contained head or tail. The data coming from the top must be checked by the head. The purpose of the tail command is to check the data starting at the bottom. The third and last element of this research is the confirmation as to whether the tail as well as head came first.

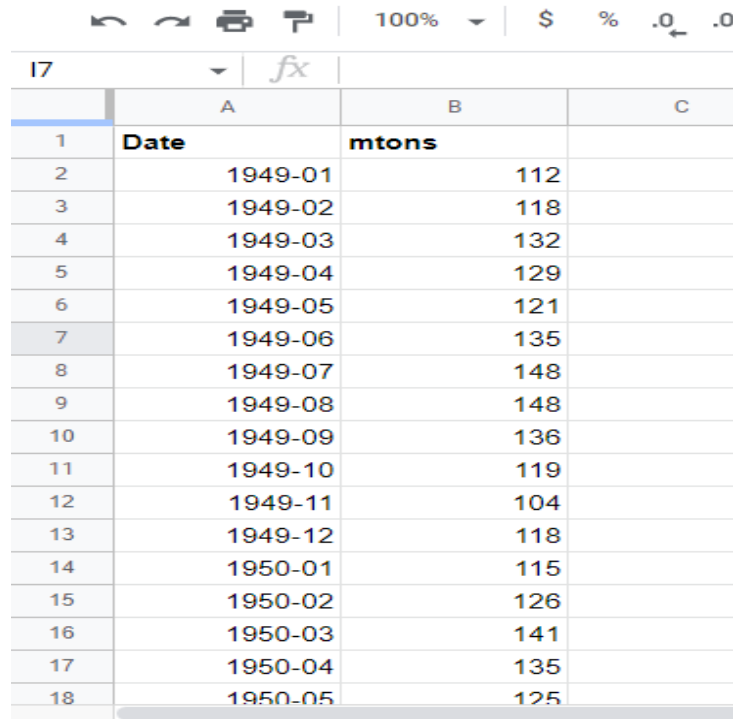
The stage four about the advanced analytics process began after the information was confirmed. Consists in calculating the rolling statistics. The mean and the standard deviation are the two essential components that are taken into consideration. The Dickey–Fuller test is to be carried out as the fifth step of this investigation. In this stage, we will investigate whether or not an autoregressive model contains a unit root by testing the "null hypothesis." The many versions of the test each have their own unique implications for the research explanation. However, the conclusion of the test might demonstrate that it is stationary or that it is trending immobile. After completing the Dickey–Fuller test, the experiment will then move on to the transformation to stationary phase. During this stage, which was led by determining an estimate of the trend by charting on a log scale.

Carry on with the experiment; the project goal entails examining the results of the Dickey – Fuller test. The crucial value needs to be somewhat near to the test data in order to achieve a high level of accuracy. In the event that the threshold also isn't fulfilled, log scaling of the data or exponential decay weighted averaging was carried out in order to get a stable time series. Python's plot function provides a visual representation of the data, making it easier to make comparisons between the log scale and the exponential decay. In the event that the outcome does not demonstrate an improvement, then the investigation will be pursued using log transformation. The p-value from the earlier Dickey Fuller test needs to be lower than the one from the more recent Dickey Fuller test.

Plotting the Auto Correlation Function (ACF) and the Partial Auto Correlation Function is the experiment's ninth stage (PACF). When $y = 0$, the X value is assessed at both plots. This research goes to phase 10, where auto regressive (AR), moving average (MA), and auto regressive integrated moving average (ARIMA) were displayed, after evaluating auto correlation and partial auto correlation function. There must be a low value for the Residual Sum of Squares (RSS). Better AR, MA, and ARIMA models are produced by low values. The experiment's last stage is to move on with the prediction and reverse the conversion of the rejected parts. The step count was changed to alter the forecast. Each level in this data corresponds to a month.

3.3 Data Collection

This study uses secondary data from Bangladesh's wheat crop to estimate growth rates and forecast area and production. The region's production information for the wheat crop was gathered from the Bangladesh Wheat and Maize Research Institute for the years 1949 to 2018.



	A	B	C
1	Date	mtons	
2	1949-01	112	
3	1949-02	118	
4	1949-03	132	
5	1949-04	129	
6	1949-05	121	
7	1949-06	135	
8	1949-07	148	
9	1949-08	148	
10	1949-09	136	
11	1949-10	119	
12	1949-11	104	
13	1949-12	118	
14	1950-01	115	
15	1950-02	126	
16	1950-03	141	
17	1950-04	135	
18	1950-05	125	

Figure 3.3: Data entry in spreadsheet

The model building and forecasting were done using data from the agricultural years 1949 to 1990. The model's validity was tested using data from 1990 to 2018

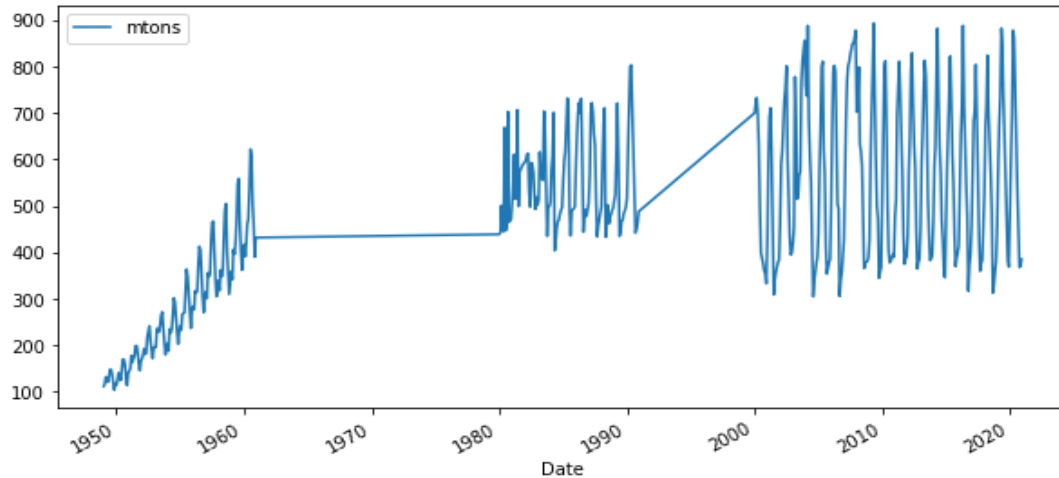


Figure 3.3.1: Visualization of Collected Data

3.4 Importance of Stationarity

A stationary time-series is an explicitly refer that doesn't rely upon that particular time. Furthermore, time-series data with patterns and/or seasonality cannot be referred to or characterized as stationary time-series data [1833]. The consecutive (hourly/daily/monthly) values in a stationary time-series do not depend on each other [1923].

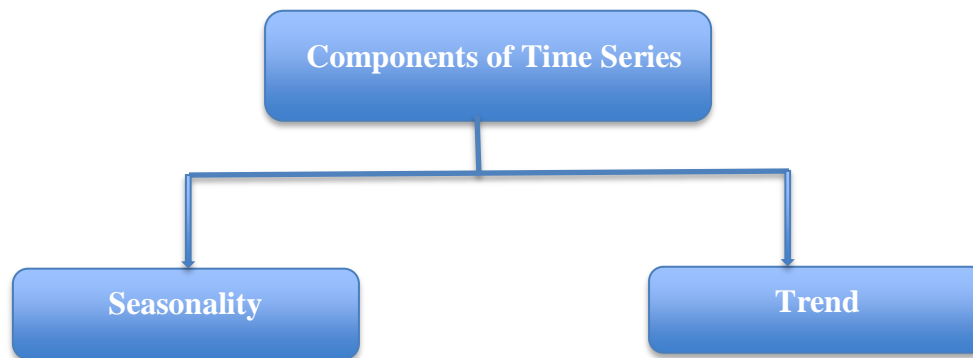


Figure 3.4: Components of time series

This explains why the data pattern is discontinuous. However, for weather forecasting, a dataset pattern must be continuous, which indicates that the sequential data must be dependent on one another. This is why non-stationary data is required for weather forecasting.

3.4.1 Trends

A trend is detected when there is a pattern in the data with regard to time. In other words, we can see a growth or decrease in the significance of a information over time. Furthermore, the trend may be broken into temporal subsections, and we can still detect trends as long as there is some form of structure. In these other words, it could be rising one time while dropping the next. However, dependent on a trigger of some kind, it could still be considered a pattern [20]. The trend might be either linear or non-linear. It is non-linear in the instance of the specified dataset.

3.4.2 Seasonality

Seasonality may be viewed as a component of a data's general pattern. Seasonality defines a single those times or even highlights its recurring areas, whereas trends monitors information more than a long time (containing several repeated periods). By the way, when it comes to weather, seasonality may be monitored across 365 days or one year. Furthermore, to see seasonality in production or any other form of data, the information must be captured in a short period of time and on a regular basis, starting from seconds [22]. However, in this study, I will just address one component, which is trend. Before proceeding with the study, it was determined whether or not the information was stationary.

3.5 Time series analysis

An assortment of information which has been compiled over the course given occasionally referred to as a time series. In time series is indeed the term given to the method that organizing quantitative information in accordance with the advancement of time. A time series is a group of readings from various points in time for a variable or combination of variables related to rice production. A time series is defined mathematically by the functional connection

Unless the mean and variance among a time series are perpetual across time, or whether the value including the covariance between time sessions varies solely on the length or interval or lag between the sample dates, instead of the clock time, then we say that perhaps the time series is stationary.

$$Y_t = f(t) \dots\dots\dots (I)$$

Where, Y_t is the value of the variable under consideration at time t .

A time series $\{Y_t, t=0, \pm 1, \pm 2, \dots\}$ is said to be stationary if it has similar statistical properties to the “time shifted” series $\{Y_{t+h}, t=0, \pm 1, \pm 2, \dots\}$ for each integer h . Simply, a time series $\{Y_t, t=0, \pm 1, \pm 2, \dots\}$ is said to be stationary time series if it is independent of time t .

A time series $\{Y_t, t=0, \pm 1, \pm 2, \dots\}$ is said to be non-stationary time series if its mean or variance or both varied over time.

For, example stock prices or exchange rates are non-stationary.

3.4.1 Auto-Covariance Function (ACVF)

Suppose, $\{y_t\}$ be a time series. Then the auto covariance function at lag k denoted by γ_k , is defined as,

$$\begin{aligned} \Gamma_k &= \text{COV}(y_t, y_{t+k}) \\ &= \frac{\sum(y_t - \mu)(y_{t+k} - \mu)}{n} \end{aligned}$$

3.4.2 Auto-Correlation Function (ACF)

Let $\{y_t\}$ be any time series, then the ACF of $\{y_t\}$ is defined as

$$\rho_k = \frac{\gamma_k}{\gamma_0} \dots \dots \dots \text{(II)}$$

Where γ_k is the covariance at lag k and γ_0 is the variance. According to our data the result are shown in figure 7&8.

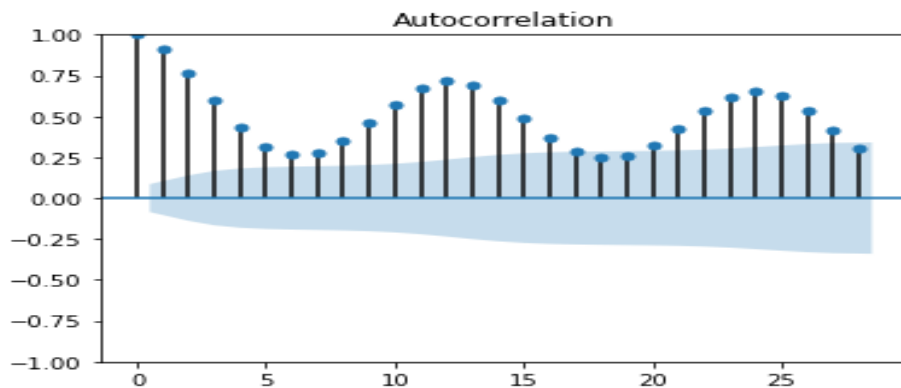


Figure 3.4.2: Autocorrelation

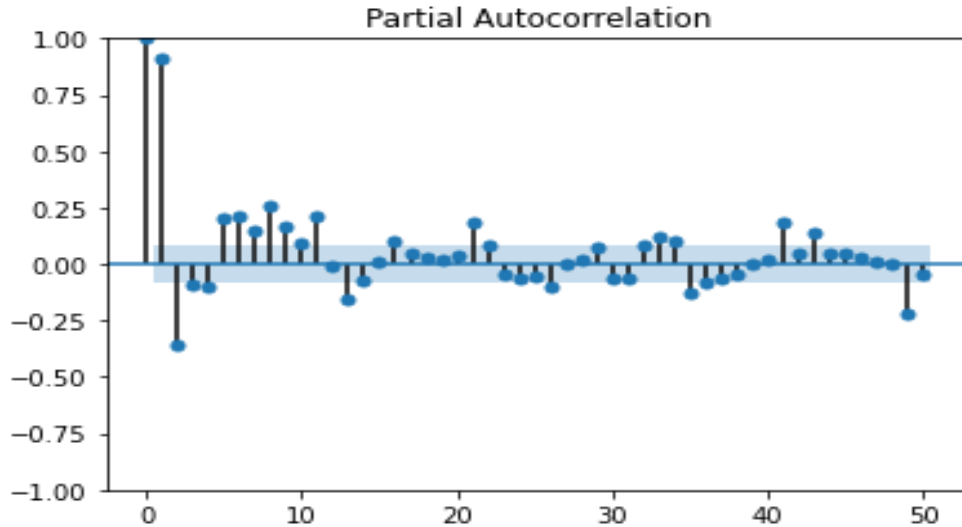


Figure 3.4.3: Partial Autocorrelation

3.5 ARIMA (Auto Regressive Integrated Moving Average) Model

The term "Auto Regressive Integrated Moving Average" (ARIMA) refers to a class of models that makes use of a time series' own previous values—more general and particularly, the exclusive lags and lagged generalization error sequence to "describe" its time - series data and make predictions about its parameter estimates. The moniker ARIMA continues to stand for "Auto Regressive Integrated Moving Average."

ARIMA models may be used to represent any time series that does not follow a seasonal pattern and does not consist entirely of random white noise. The definition of an ARIMA model may be broken down into three parts: p , d , and q . Where p represents the order about the AR term, q represents equal order about the MA term, and d is often number of differencing that are necessary to make the time series stable. When seasonality patterns are present in a time series, the time series is referred to as SARIMA, which is an abbreviation for "Seasonal ARIMA." Seasonal words should be introduced when this occurs.

The auto-regressive integrated moving average (ARIMA) model consists using linear regression which just uses the model's own lags as predictors. When the variables vary independently of one another and not connected to one another, linear regression models perform at their peak. Differentiating from many other things represents the most common approach. To put it another way, take the current value and deduct the value from when it was before. Perhaps there are

occasions when many than one differencing is necessary, and this is determined by the intricacy of the series.

As a consequence, this has, the value of d represents the minimum number of differentiations that are necessary to render the series stationary. Whereas if time series has been stationary, then d will equal 0; otherwise, it will not.

The letter 'p' comes first in the term 'auto-regressive,' which is an acronym. It indicates the total number of Y lags that will be used as predictors in the analysis. The letter 'q' indicates the sequence in which the term 'Moving Average' (MA) is presented. It is the total amount of mistakes in the lag forecast that have to be incorporated into the ARIMA model.

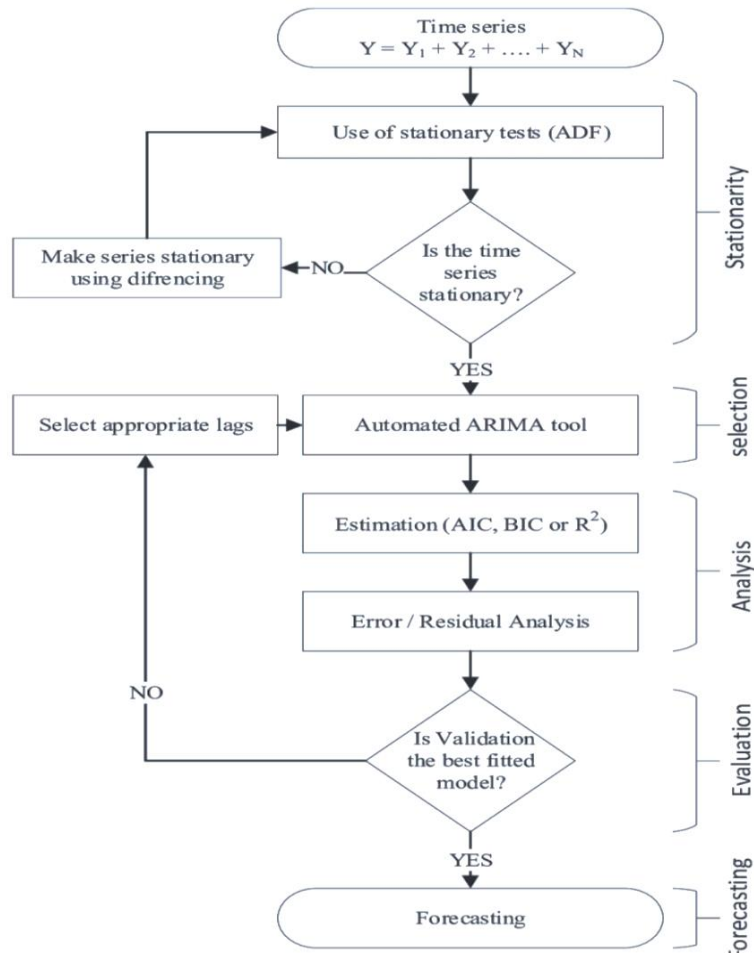


Figure 3.5: Working Process of ARIMA

3.5.1 AR and MA models

A pure Auto Regressive (AR alone) model is one in which Y_t is only determined by its own lags. That is, Y_t is a function of the 'Yt lags.'

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + \epsilon_t$$

Where Y_{t-1} is the series' lag1, β_1 is the lag1 coefficient predicted by the model, and α is the intercept term estimated by the model.

Similarly, a pure Moving Average (MA alone) model is one in which Y_t is determined solely by the lagged forecast errors.

$$Y_t = \alpha + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \dots + \phi_q \epsilon_{t-q}$$

Where the error terms are the errors of the respective autoregressive models of the lags. The errors ϵ_t and ϵ_{t-1} result from the following equations:

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_0 Y_0 + \epsilon_t$$

$$Y_{t-1} = \beta_1 Y_{t-2} + \beta_2 Y_{t-3} + \dots + \beta_0 Y_0 + \epsilon_{t-1}$$

An ARIMA model is one in which the time series is differenced at least once to make it stationary and the AR and MA terms are combined. As a result, the equation is:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \dots + \beta_p Y_{t-p} + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \dots + \phi_q \epsilon_{t-q}$$

Predicted $Y_t = \text{Constant} + Y \text{ Delays (up to } p \text{ lags)} + \text{Linear Combination of Lagged Forecast Errors (upto } q \text{ lags)}$. The goal is thus to determine the values of p , d , and q .

3.5.2 SARIMA

ARIMA, which stands to Autoregressive Integrated Moving Average, is one of the most used forecasting models for univariate time series data. Although this method can handle statistical model towards a forecast, something can handle trending data. SARIMA is an extension of ARIMA that permits significant modeling of the constant term of a series.

The Autoregressive Integrated Moving Average (ARIMA) is a method for predicting univariate time series data. As its name suggests, it supports both autoregressive and moving average components. The integrated feature relates to differencing, which permits the method to accommodate trending time series data. ARIMA is limited in that it cannot process seasonal data. This time series demonstrates a cyclical trend. ARIMA predicts data that is either not seasonally dependent or has had the seasonal component removed, such as data that has been seasonally adjusted utilizing techniques such as seasonal differencing.

SARIMA (Seasonal Autoregressive Integrated Moving Average) or Seasonal ARIMA is an ARIMA version that explicitly accepts seasonal components into multivariate time series data. It has three new hyperparameters to characterize the autoregression (AR), differencing (I), and moving average (MA) for the seasonal component of a series, as well as a parameter for the seasonality period.

There are three trend elements that need to be configured. They are identical to the ARIMA model, specifically:

- p: trend autoregression order.
- d: Order of trend difference.
- q: trend moving average order.

There seem to be four seasonal components that must be configured that are not part of ARIMA; they are as follows:

We used discretionary wheat creation data for Bangladesh from 1949 to 2018 to find an example of wheat creation in Bangladesh. The information was gathered from numerous conveyances of Bangladesh estimating authority. Box and Jenkins devised the Box-Jenkins approach in 1970 to determine what happened later on, whether the record climbs or falls. It is the most appropriate model determination technique for forecasting time arrangement variables.

The Box-Jenkins method is useful only when the variable confirms a few suspicions. As a matter of first importance, the variable should be fixed and there should be no abnormality; however, if the variable rejected the notion or any model as evidenced by scale or region, the primary aim is to make it fixed for using the Box-Jenkins model. In practice, most factors are non-fixed, therefore to convert non-fixed elements into fixed components, we gain the qualification of elements, and the new factors are known as consolidated variables.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

This model is a generalization of the ARMA model that uses the integrated moving average of past data to predict future outcomes. These two models are commonly used for predicting new observations in time series. To illustrate overall power among predicting parameters related to and from other varying variables, ARIMA is a type of statistical study. Throughout order to predict how a time series will change overtime, it model looks at the differences between individual numbers in the series rather than their absolute values. When there are indications of non-stationarity in the data, ARIMA models are applied. In time series analysis, non-stationary information is always converted to stationary data.

4.2 Stationary Test:

According to collected data:

Table 1: Stationary Test Result

Name	Value
ADF	112
P-value	118
Num of lags	132
Number of observations used for ADF regression and critical values calculation	129
Critical-value	1%: -3.4432119442564324 5%: -2.8640418343195267 10%: -2.568770059171598

In this chapter, we determine the production trend line and then determine whether or not the data is stationary. If the data is nonstationary, we make it stationary by collecting differences and eventually estimating wheat production using ARIMA time series forecasting and fitting the ARIMA model.

In ascertain if a given time assumption holds, statisticians often rely to anything other than the Augmented Dickey Fuller test (ADF Test). It is among the most common parametric analyses used to determine whether or not a data is normal.

4.3 Model Identification and Parameter Estimation

In order to determine how well a model matches its input data, statisticians employ the Akaike information criterion (AIC). In stats, using Akaike Information Criterion (AIC) is employed to compare potential models to choose the model that best fits the data. The Akaike Information Criterion is computed by taking the expectation - maximization estimates of both the structure and divide by the total set of individual variable in the model (how well the model reproduces the data). The AIC states that somehow the optimal model is the one that adequately describes the data while using the fewest possible explanatory variables.

Table 2: Model Identification

ARIMA(2,1,2)(0,0,0)[0] intercept	: AIC=6062.604, Time=0.60 sec
ARIMA(0,1,0)(0,0,0)[0] intercept	: AIC=6108.180, Time=0.02 sec
ARIMA(1,1,0)(0,0,0)[0] intercept	: AIC=6058.874, Time=0.10 sec
ARIMA(0,1,1)(0,0,0)[0] intercept	: AIC=6060.703, Time=0.13 sec
ARIMA(0,1,0)(0,0,0)[0]	: AIC=6106.203, Time=0.02 sec
ARIMA(2,1,0)(0,0,0)[0] intercept	: AIC=6060.818, Time=0.12 sec
ARIMA(1,1,1)(0,0,0)[0] intercept	: AIC=6060.814, Time=0.19 sec
ARIMA(2,1,1)(0,0,0)[0] intercept	: AIC=5970.388, Time=0.36 sec
ARIMA(3,1,1)(0,0,0)[0] intercept	: AIC=6061.519, Time=0.37 sec
ARIMA(1,1,2)(0,0,0)[0] intercept	: AIC=6062.788, Time=0.26 sec
ARIMA(3,1,0)(0,0,0)[0] intercept	: AIC=6062.820, Time=0.07 sec
ARIMA(3,1,2)(0,0,0)[0] intercept	: AIC=6045.772, Time=0.78 sec
ARIMA(2,1,1)(0,0,0)[0]	: AIC=5971.781, Time=0.20 sec

Best model: ARIMA (2,1,1) (0,0,0) [0]

Total fit time: 3.223 seconds

According to the best model (2,1,1), after fitting the model this is the result:

```

model = ARIMA(df.mtons, order=(2,1,1))
model=model.fit()
model.summary()

```

SARIMAX Results

Dep. Variable:	mtons	No. Observations:	528
Model:	ARIMA(2, 1, 1)	Log Likelihood	-2981.890
Date:	Fri, 26 Aug 2022	AIC	5971.781
Time:	15:56:07	BIC	5988.849
Sample:	0	HQIC	5978.463
	- 528		

Figure 4.3: Analysis Result

4.3.1 Covariance

Table 3: Results of Covariance

	coef	std err	Z	P> z	[0.025	0.975]	coef
intercept	0.2183	0.117	1.866	0.062	-0.011	0.448	0.2183
ar.L1	1.1705	0.033	35.175	0.000	1.105	1.236	1.1705
ar.L2	-0.4359	0.034	-12.706	0.000	-0.503	-0.369	-0.4359
ma.L1	-0.9658	0.012	-80.139	0.000	-0.989	-0.942	-0.9658
sigma2	4764.9695	214.968	22.166	0.000	4343.640	5186.299	

Covariance Type: opg						
	coef	std err	z	P> z	[0.025	0.975]
intercept	0.2183	0.117	1.866	0.062	-0.011	0.448
ar.L1	1.1705	0.033	35.175	0.000	1.105	1.236
ar.L2	-0.4359	0.034	-12.706	0.000	-0.503	-0.369
ma.L1	-0.9658	0.012	-80.139	0.000	-0.989	-0.942
sigma2	4764.9695	214.968	22.166	0.000	4343.640	5186.299
Ljung-Box (L1) (Q):	4.97	Jarque-Bera (JB):	193.72			
Prob(Q):	0.03	Prob(JB):	0.00			
Heteroskedasticity (H):	1.55	Skew:	-0.30			
Prob(H) (two-sided):	0.00	Kurtosis:	5.91			

The ARIMA (4,1,3) process is defined,

$$Y_t = \phi_1 Y_{t-1} + Z_t \dots\dots\dots (III)$$

Where $\{Z_t\}$ follows White Noise $(0, \sigma^2)$ and ϕ constant.

4.4 Model Performance and Evaluation Matric

Model performance is an evaluation of a model's capacity to perform a task accurately not just using training data but also in real-time with runtime data when the model is deployed.

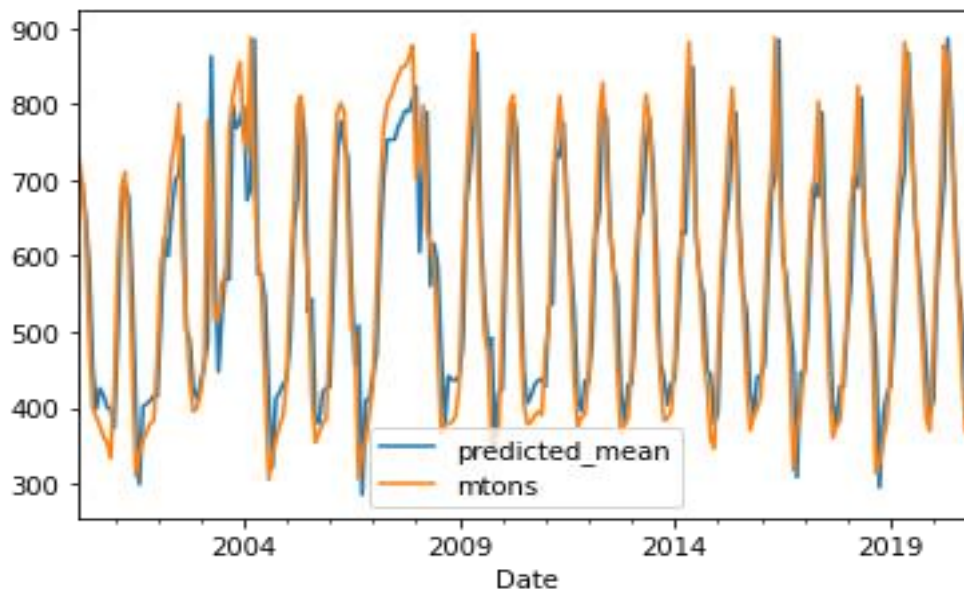


Figure 4.4: Original Vs Predicted Result

In statistics, the term "mean squared error" (MSE) refers to an average value calculated by squaring the disparity between the observed and predicted values. So here the accuracy is 88%.

```

▶ test['mtons'].mean()
↳ 565.536

[ ] from sklearn.metrics import mean_squared_error
    from math import sqrt
    rmse=sqrt(mean_squared_error(pred,test['mtons']))
    print(rmse)

75.44310290547156

```

Figure 4.4.1: Result of MSE

4.5 Analysis of Forecasting Results

Table 4: Forecasting Results

2019	2020	2021	2022	2023	2024	2025	2026	2027	2028
478	584	550	510	601	590	517	494	680	582

From 1949 to 2018 we get the wheat creation information in Bangladesh. Presently we get the investigation anticipating results for forwarding 10 years.

CHAPTER 5

CONCLUSION

6.1 Conclusion

According to the findings of the study, Bangladesh is seeing an upward trend pattern in terms of the productivity and yield of its wheat crops. Wheat output and yield have both shown positive growth rates across all of the subperiods. The first subperiod, 2000-2018, showed the largest rise in production and yield, whereas the second subperiod, 1990-1999, saw the most expansion in area. The first subperiod also saw the biggest growth in output and yield. When it comes to modeling and forecasting wheat production, it was found that ARIMA (2,1,1) with drift performed the best out of all of the ARIMA models that were tested with 88% accuracy. Utilizing time series analysis allows us to arrive at conclusions on the future of wheat cultivation in Bangladesh. In order to anticipate the output of wheat in Bangladesh for the next 10 years, we will be using the ARIMA model.

The comparison plot between the model's predicted values and the actual values reveals that the model is an accurate representation of the data. In conclusion, one of the takeaways from this research is that in order to maximize wheat production, one must choose the most effective method of growing. Farmers are required to plant crops using the highest quality seeds, as well as an adequate amount of fertilizer and weed killer. Wheat may be grown in huge quantities provided the farmer has the necessary skills. The numerous costs should not escape the attention of farmers. In order to optimize output and profit, costs need to be cut down significantly. When it comes to productivity, farmers need to pay attention to both temperature and rainfall.

6.2 Future Work

In sum, ARIMA is an effective statistical method for handling large datasets. While there are several algorithms that may be used for this purpose, ARIMA provides a clearer path to data prediction and a deeper knowledge of the process. The study's ultimate goal is to employ an algorithm for data prediction of manufacturer-produced rejects. Humans are emotional creatures, and they can easily grow weary from doing the same actions over and over again. As a result, defective goods could be manufactured. There is a lot of room for growth and improvement in this

project using various machine learning algorithms. A digital record keeping system and more data from other regions might be beneficial to this endeavor. The software can be connected to agriculture servers to deliver instantaneous updates. Classifying data with more variables, such as urban/rural location, hygiene, and food nutrition information, can lead to more precise forecasts.

REFERENCES

- [1] 'Historic 7 March to observe in Rangpur division', The Asian Age, 2016. Available: <https://dailyasianage.com/news/12749/historic-7-march-to-observe-in-rangpur-division>
- [2] 'Bangladesh Wheat Imports by Year', Bangladesh Wheat Imports by Year (1000 MT), Available: <https://www.indexmundi.com/agriculture/?country=bd&commodity=wheat&graph=imports>
- [3] P. Mishra, P.K. Sahu, B.S. Dhekale, K.P. vishwajith, Modeling and forecasting of wheat in India and their yield Sustainability, J. Econ. Dev. 11 (3) (2015) 637–647. Indian.
- [4] Anonymous, ICAR – Vision 2020, Indian Council of Agricultural Research, New Delhi, India, 1999.
- [5] D. Malik, D. Singh, Dynamics of production, processing and export of wheat in India, J. Food . Secur. 1 (2010) 1–12, 1.
- [6] P.J. Brockwell, R.A. Davis, Introduction to Time Series and Forecasting, 2nd.ed., Springer Verlag, 2002
- [7] R.P. Singh, S.K. Das, R.V.M. Bhaskar, N.M. Reddy, Towards Sustainable Dryland Agricultural Practices, Technical Bulletin, Central Institute for Dryland Agriculture, Hyderabad, India, 1990, p. 106.
- [8] P. Mishra, M.G. Al Khatib, I. Sardar, et al., Modeling and forecasting of sugarcane production in India, Sugar Tech (2021) 1–8, <https://doi.org/10.1007/s12355-02101004-3>
- [9] R. Chand, S.S. Raju, Instability in Indian agriculture during different phases of technology and policy, Indian J. Agric. Econ. 21 (2) (2009) 283–288.
- [10] J.K. Lynam, in: P. Goldsworthy, F.W.T.) Penning de Vries (Eds.), Opportunities, Use, and Transfer of Systems Research Methods in Agriculture in Developing Countries, Kluwer Academic Publishers, Dordrecht, Netherlands, 1994, pp. 3–28.
- [11] H. Tyagi, S. Suran, and V. Pattanaik, "Weather-temperature pattern prediction and anomaly identification using artificial neural network," International Journal of Computer Applications, vol. 975, p. 8887, 2016.
- [12] oTexts, Forecasting: Principles and practice (2nd edition). [Online]. Available: <https://otexts.com/fpp2/stationarity.html>.
- [13] E. Cadenas, W. Rivera, Wind speed forecasting in three different regions of Mexico, using a hybrid ARIMA–ANN model, Renew. Energy 35 (12) (2010) 272–273.
- [14] L.A. Díaz-Robles, J.C. Ortega, J.S. Fu, G.D. Reed, J.C. Chow, J.G. Watson, J. A. Moncada-Herrera, A hybrid ARIMA and artificial neural networks to forecast particulate matter in urban areas: the case of Temuco, Chile, Atmos. Environ. 42 (35) (2008) 8331–8340.
- [15] D.O. Faruk, A hybrid neural network and ARIMA model for water quality time series prediction, Eng. Appl. Artif. Intell. 23 (4) (2010) 586–594.
- [16] P. Mishra, A. Matuka, M.S.A. Abotaleb, et al., Modeling and forecasting of milk production in the SAARC countries and China, Model. Earth Syst. Environ. (2021) 1–13.
- [17] P. Mishra, P.K. Sahu, B.S. Dhekale, K.P. vishwajith, Modeling and forecasting of wheat in India and their yield Sustainability, J. Econ. Dev. 11 (3) (2015) 637–647

- [18] M. Ray, A. V. Rai, Ramasubramanian, K.N. Singh, ARIMA-WNN hybrid model for forecasting wheat yield time series data, *J. Indian Soc. Agric. Stat.* 70 (1) (2016) 63–70.
- [19] N.A. Saeed, M. Saeed, Zakria, T.M. Bajwa, Forecasting of wheat production in Pakistan using ARIMA models, *International Journal of Agricultural Biology* 2 (4) (2000) 352–353.
- [20] Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. *Forecasting Methods and Applications* (3rd ed.) 1998; New York: John Wiley & Sons, Inc.
- [21] T. S. A. Toppr, Components of time series, 2021. [Online]. Available: <https://www.toppr.com/guides/business-mathematics-and-statistics/time-seriesanalysis/components-of-time-series/>.

Wheat Production Forecasting in Bangladesh

ORIGINALITY REPORT

13% SIMILARITY INDEX	11% INTERNET SOURCES	4% PUBLICATIONS	11% STUDENT PAPERS
--------------------------------	--------------------------------	---------------------------	------------------------------

PRIMARY SOURCES

1	repositorio.unal.edu.co Internet Source	2%
2	Submitted to Daffodil International University Student Paper	2%
3	machinelearningmastery.com Internet Source	1%
4	Submitted to Associatie K.U.Leuven Student Paper	1%
5	www.machinelearningplus.com Internet Source	1%
6	Submitted to University of Newcastle upon Tyne Student Paper	1%
7	techniumscience.com Internet Source	1%
8	Submitted to Liverpool John Moores University Student Paper	1%
9	Submitted to University of Sydney	