AGRICULTURAL TIME SERIES ANALYSIS FOR DIFFERENT CROPS ACROSS DIFFERENT STATES OF INDIA

Ankita Pal & Nilanjan Barik

Registration Number : 23414270014 & 23414110021

Class Roll Number : STAT 014 & STAT 021

PRESIDENCY UNIVERSITY
KOLKATA

Department of Statistics

Presidency University

Kolkata

Acknowledgement

We would like to express our profound gratitude to Prof. Suman Guha, our supervisor for his contributions and efforts he provided throughout the year to completion of this project titled "**Agricultural Time Series Analysis For Different Crops across Different States of India**".

We would like to express our special thanks to Prof. Radhakanta Das, HOD, Department Of Statistics, Presidency University, Kolkata, for his time and continuous support throughout the journey.

We are indebted to Kaggle for providing access to the Crop Production Data over India throughout the year. Their generous support has been pivotal in enabling the statistical analysis and exploration conducted in this project.

Finally, we would like to extend our sincere appreciation to our family, friends, and loved ones who have provided unwavering support, encouragement, and understanding throughout this project. Their belief in us and their willingness to lend a helping hand have been a constant source of motivation.

Ankita Pal & Nilanjan Barik

Reg. No : 23414270014 & 23414110021

CONTENTS

Sl No	Subjects	Page No
1.	*Introduction*	*1-4*
1.1	*Background of Research*	*1*
1.2	*Motivation Behind the study*	*1-2*
1.3	*Outline of Research*	*2*
1.4	*Literature Review*	*2-3*
1.5	*Data Description*	*3-4*
2.	*Method of Analysis*	*4-32*
2.1	*Exploratory Analysis*	*4-9*
2.2	*Spatial Autocorrelation*	*9-12*
2.3	*Forecasting Using ARIMA*	*13-31*
2.3	*Forecasting Using ARIMAX*	*31-32*
3	*Findings and conclusion*	*32-33*
4	*References.*	*33*

1. INTRODUCTION

1.1 BACKGROUND OF RESEARCH

The history of **agriculture in India** dates back to the **Neolithic Period**. India ranks second worldwide in farm outputs. As per the Indian economic survey 2020 -21, agriculture employed more than 50% of the Indian workforce and contributed 20.2% to the country's GDP. Agriculture has long been the backbone of India's economy, contributing significantly to the nation's GDP and serving as the primary source of livelihood for a large segment of the population. With diverse agro-climatic conditions, India grows a wide variety of crops across its states and across three main cropping seasons: Kharif, Rabi, and Zaid. However, the production of these crops is subject to fluctuations due to a multitude of factors including seasonal variability, rainfall patterns, pest outbreaks, and changing agricultural practices. Accurately predicting crop yield is essential for ensuring food security, planning procurement, managing supply chains, and formulating effective agricultural policies.

1.2 MOTIVATION BEHIND THE STUDY:

India's agriculture sector is a critical component of its economy, yet it remains vulnerable to unpredictable climatic patterns, seasonal fluctuations, and inconsistent crop yields. These uncertainties pose significant challenges not only to farmers but also to policymakers, agricultural planners, and food supply chains. Despite the availability of vast historical crop production data across Indian states, the full potential of data-driven techniques to predict future yields remains underutilized.

The motivation behind choosing this topic stems from the pressing need to harness **data analytics and statistical modeling** to support **agricultural forecasting and planning**. With increasing climate variability and the growing demand for food security, there is a strong imperative to develop predictive models that can inform timely decisions regarding resource allocation, procurement strategies, and crop management practices.

We were particularly driven by the opportunity to apply **time series analysis and ARIMA modeling** to real-world agricultural datasets, enabling us to convert historical crop yield data into actionable insights. These methods are not only well-established in forecasting but also offer interpretability and adaptability, which are crucial in agricultural contexts.

By focusing on a specific crop and state, we aim to provide a **focused yet scalable approach** to yield prediction that can be adapted to different crops, regions, and climatic zones. This project also aligns with national and global efforts to strengthen climate-resilient agriculture through the use of technology and data science.

Ultimately, our motivation is to contribute to a **more predictive and proactive agricultural system**, where stakeholders can make informed decisions based on reliable forecasts leading to better yield management, reduced risk, and improved food security outcomes.

1.3 OUTLINE OF RESEARCH:

We first start our analysis using the Exploratory Data Analysis to visualize and examine the key features of the yield of some of the major crops of the States of India through the year 1997 to 2019. We have gained a rough sense about the key trends , patterns, seasonal effect of yield of some crops .

Now we are going to focus the time series forecasting particularly ARIMA developed for some major crops in each state. These models are used to predict future yields, enabling more informed planning and decision making in the agricultural sector.

Moving from ARIMA model we are going to forecast the yield with some relatable variable using ARIMAX model. This model solely predict the yield values that depends on some other variables like annual rainfall, pesticides or fertilizers.

1.4 LITERATURE REVIEW:

The backbone of India's economy is Agriculture. There is an increased requirement to predict the future crop yield to match the crop demands. Farmers want to know which crop to plant and approximate yield in advance. However, unpredictable rainfall trends, seasonal production trends, and multiple climatic aspects make it challenging to recommend crops and predict yield.

Keeping mind of this increasing demand of future prediction of crop yield and the effect of parameters that associated with crop production and yield like rainfall, pesticides, soil temperature etc; many research organisations, Indian Government, State Government machinaries, independent reasearchers are working on this field to give the real time accurate predictions not only on crop production but also their association with other aggricultural parameters that mentioned above.

- ***A Model for Prediction of Crop Yield (International Journal of Computational Intelligence***

and Informatics, Vol. 6: No. 4, March 2017) https://www.periyaruniversity.ac.in/ijcii/issue/Vol6No4Mar2017/M5_PID0370.pdfhttps://www.periyaruniversity.ac.in/ijcii/issue/Vol6No4Mar2017/M5_PID0370.pdf is such a reaserch paper that shows us that the data mining is one of the emerging research field in crop yield analysis.

- ***USING THE ARIMA MODELS TO PREDICT WHEAT CROP PRODUCTION IN IRAQ (*Int. J. Agricult. Stat. Sci. Vol. 16, No. 1, pp. 121-127, 2020 *)*** https://connectjournals.com/file_full_text/3159801H_121-127.pdf

this research paper also employ the ARIMA model for predicting the wheat production in Iraq.

- ***FORECASTING FOR AGRICULTURAL PRODUCTION USING ARIMA MODEL (Palarch's Journal Of Archaeology Of Egypt/Egyptology 18(7).ISSN 1567-214x)*** https://archives.palarch.nl/index.php/jae/article/download/5116/5043/9832

This paper also use the ARIMA model for paddy production forecast from South India . The fitted ARIMA models that are used for Andhra Pradesh, Karnataka, Kerala and Tamil Nadu are ARIMA(0,1,1), ARIMA(0,1,1), ARIMA(0,1,2), ARIMA(0,1,1). The models are examine through computing the various parameters estimate measures like BIC, MSE, etc.

1.5 DATA DESCRIPTION:

For investigating the Agricultural crop yield in Indian states , Data are collected from Kaggle.

This dataset includes agricultural data for multiple crops cultivated across various states in India from the year **1997 till 2020**. The dataset provides crucial features related to crop yield , including crop types, crop years, cropping seasons, states, areas under cultivation, production quantities, annual rainfall, fertilizer usage and pesticide usage.

The columns description of the dataset are given as follows:

1. **Crop**: The name of the crop cultivated.
2. **Crop_Year**: The year in which the crop was grown.
3. **Season**: The specific cropping season (e.g., Kharif, Rabi, Whole Year).
4. **State**: The Indian state where the crop was cultivated.
5. **Area**: The total land area (in hectares) under cultivation for the specific crop.
6. **Production**: The quantity of crop production (in metric tons).
7. **Annual_Rainfall**: The annual rainfall received in the crop-growing region (in mm).
8. **Fertilizer**: The total amount of fertilizer used for the crop (in kilograms).
9. **Pesticide**: The total amount of pesticide used for the crop (in kilograms).
10. **Yield**: The calculated crop yield (production per unit area).

Courtesy: Kaggle

https://www.kaggle.com/datasets/akshatgupta7/crop-yield-in-indian-states-dataset

snapshot of the dataset:

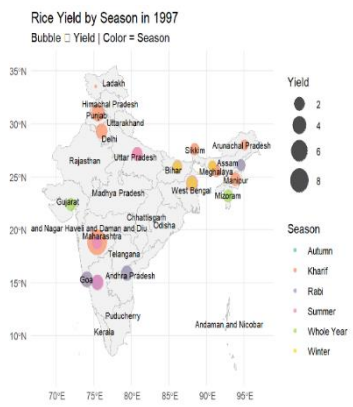
	A	B	C	D	E	F	G	H	I	J	K
1	***Crop***	***Crop_Year***	***Season***	***State***	***Area***	***Production***	***Annual_Rainfall***	***Fertilizer***	***Pesticide***	***Yield***	
2	Arecanut	1997	Whole Yea	Assam	73814	56708	2051.4	7024878	22882.34	0.796087	
3	Arhar/Tur	1997	Kharif	Assam	6637	4685	2051.4	631643.3	2057.47	0.710435	
4	Castor see	1997	Kharif	Assam	796	22	2051.4	75755.32	246.76	0.238333	
5	Coconut	1997	Whole Yea	Assam	19656	126905000	2051.4	1870662	6093.36	5238.052	
6	Cotton(lint	1997	Kharif	Assam	1739	794	2051.4	165500.6	539.09	0.420909	
7	Dry chillies	1997	Whole Yea	Assam	13587	9073	2051.4	1293075	4211.97	0.643636	
8	Gram	1997	Rabi	Assam	2979	1507	2051.4	283511.4	923.49	0.465455	
9	Jute	1997	Kharif	Assam	94520	904095	2051.4	8995468	29301.2	9.919565	
10	Linseed	1997	Rabi	Assam	10098	5158	2051.4	961026.7	3130.38	0.461364	
11	Maize	1997	Kharif	Assam	19216	14721	2051.4	1828787	5956.96	0.615652	
12	Mesta	1997	Kharif	Assam	5915	29003	2051.4	562930.6	1833.65	4.568947	
13	Niger seed	1997	Whole Yea	Assam	9914	5076	2051.4	943515.4	3073.34	0.482353	
14	Onion	1997	Whole Yea	Assam	7832	17943	2051.4	745371.4	2427.92	2.342609	
15	Other Rab	1997	Rabi	Assam	108297	58272	2051.4	10306625	33572.07	0.52087	
16	Potato	1997	Whole Yea	Assam	75259	671871	2051.4	7162399	23330.29	7.561304	
17	Rapeseed	1997	Rabi	Assam	279292	154772	2051.4	26580220	86580.52	0.554783	
18	Rice	1997	Autumn	Assam	607358	398311	2051.4	57802261	188281	0.78087	
19	Rice	1997	Summer	Assam	174974	209623	2051.4	16652276	54241.94	1.060435	
20	Rice	1997	Winter	Assam	1743321	1647296	2051.4	1.66E+08	540429.5	0.941304	
21	Sesamum	1997	Whole Yea	Assam	15765	8257	2051.4	1500355	4887.15	0.487391	
22	Small mille	1997	Kharif	Assam	10490	5391	2051.4	998333.3	3251.9	0.473	
23	Sugarcane	1997	Kharif	Assam	31318	1287451	2051.4	2980534	9708.58	41.89696	
24	Sweet pot	1997	Whole Yea	Assam	9380	32618	2051.4	892694.6	2907.8	3.440435	
25	Tapioca	1997	Whole Yea	Assam	2465	11728	2051.4	234594.1	764.15	4.418261	
26	Tobacco	1997	Whole Yea	Assam	433	26	2051.4	41208.61	134.23	0.38	
27	Turmeric	1997	Whole Yea	Assam	10071	6974	2051.4	958457.1	3122.01	0.67	
28	Wheat	1997	Rabi	Assam	84698	110054	2051.4	8060709	26256.38	1.259524	
29	Arecanut	1997	Whole Yea	Karnataka	93100	133342	1266.7	8860327	28861	1.293571	

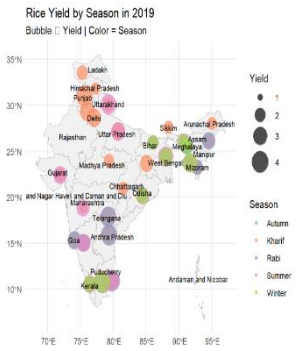
crop_yield

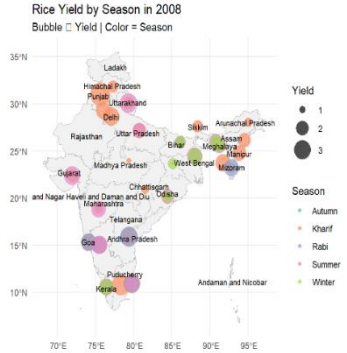
2. METHOD OF ANALYSIS

Initially We focuses on West Bengal . In depth we explore the major crops of the state and analyse their production.

2.1 EXPLORATORY DATA ANALYSIS







2.1.1 Analysis for Rice

- **Bubble size** shows the **yield** in that state.
- **Bubble color** identifies the **season** in which rice was cultivated.

- Selected three reference years: **1997**, **2008**, and **2019**.

Rice is a staple crop in India, grown across multiple seasons like **Kharif**, **Rabi**, **Autumn**, **Summer**, **Winter**, and sometimes throughout the **Whole Year**. But not all states grow rice in all seasons.

- **In 1997 :**

Goa, **Andhra Pradesh**, and **Uttar Pradesh** show **"Rabi " season yields** .

Bihar and **Manipur** show yield in "**Winter"** and "**Kharif"** respectively.

Here we can observe mostly single-season rice cultivation per state.

- **In 2008 :**

A significant **increase in multiseason cultivation**.

Goa, **Kerala**, **Andhra Pradesh**, **Maharashtra**, etc. grow rice in "**Summer"** and "**Rabi"**.

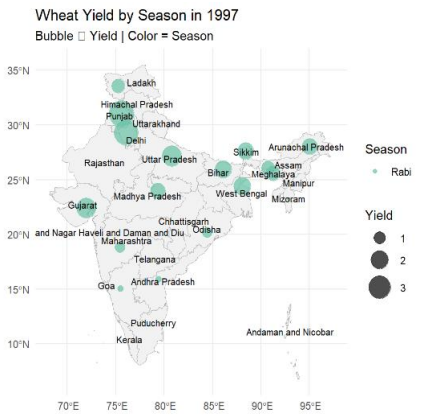
- In **2019**

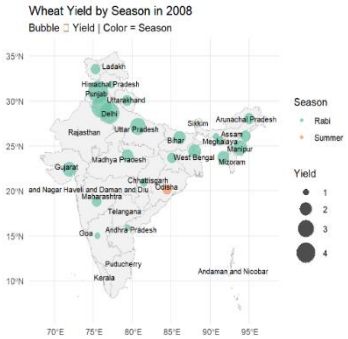
States like **West Bengal**, **Andhra Pradesh**, and **Meghalaya** show **Winter** and **Whole Year** cultivation indicating extended agricultural cycles.

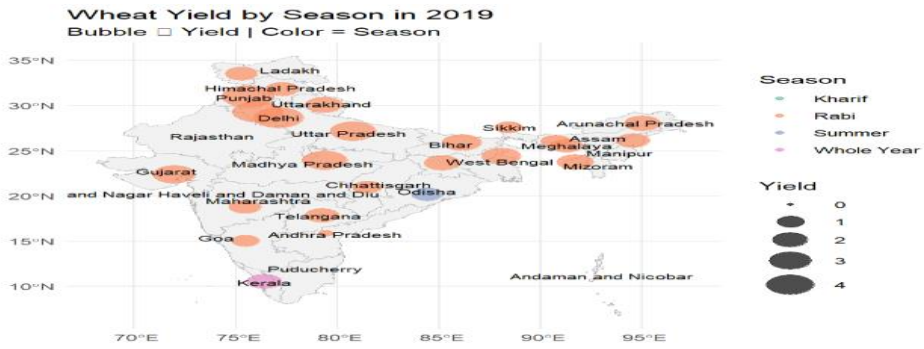
Diversity in season-wise cultivation is more visible than earlier years.

Over the years, India's rice cultivation has expanded across seasons in different states. In **1997**, rice farming was primarily limited to traditional seasons like "**Kharif"** and "**Rabi"**. By **2008**, we started noticing **multi-seasonal farming**, although yields were relatively low. Fast forward to **2019**, states adapted to various agro-climatic techniques, leading to rice cultivation in almost all seasons including **Winter** and even **Whole Year** cycles in some regions.

2.1.2 Analysis of Wheat







- **In 1997 :**

Only "**Rabi"** season is recorded.

Delhi, Uttarakhand, Bihar show relatively higher yields.

Arunachal Pradesh, **West Bengal**, and **Gujarat** have smaller yields.

Yield is limited mostly to North and Eastern India.

- **In 2008 :**

Odisha grows wheat in "**Summer"** season .

Yields have **increased** .

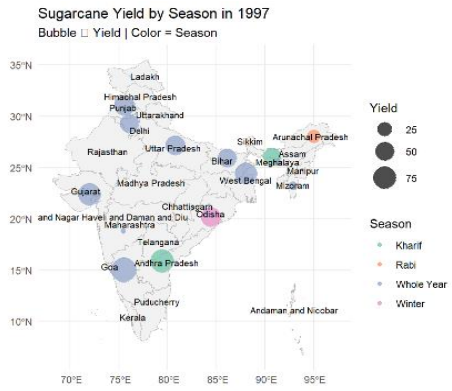
- In **2019**

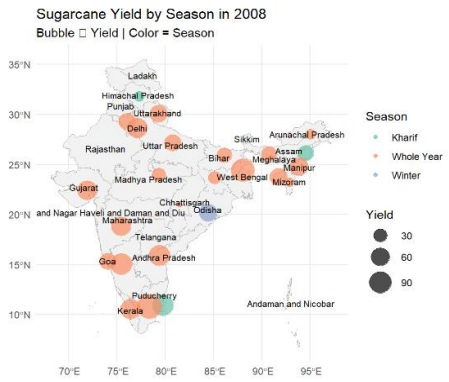
Wheat cultivation is seen **all across India**, including **Southern states** (Andhra Pradesh, Kerala, Telangana).

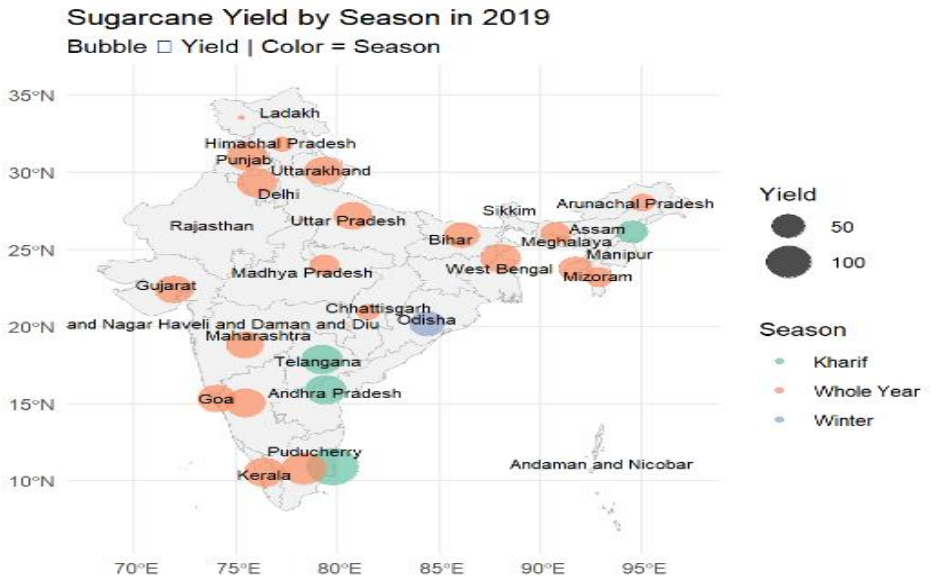
Kerala produces wheat over the **Whole Year**.

Wheat cultivation **expanded geographically and seasonally** between 1997 and 2019.

2.1.3 Analysis of Sugarcane







- **In 1997**

Many states like Maharashtra, Bihar, Andhra Pradesh, and Telangana cultivated sugarcane for the whole year.

Odisha grew it in the Winter season.

Yield levels were moderate overall .

- **In 2008**

Whole Year cultivation became widespread.

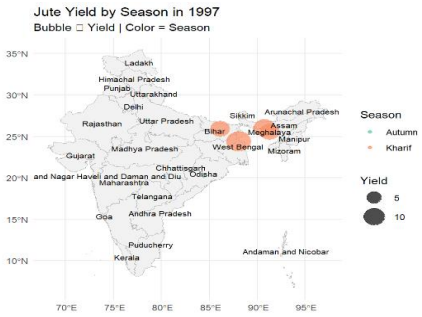
Few states like Kerala and Puducherry still practiced "Kharif" cultivation.

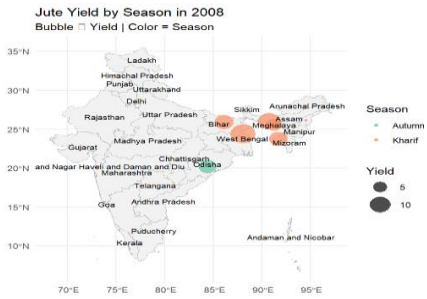
- **In 2019**

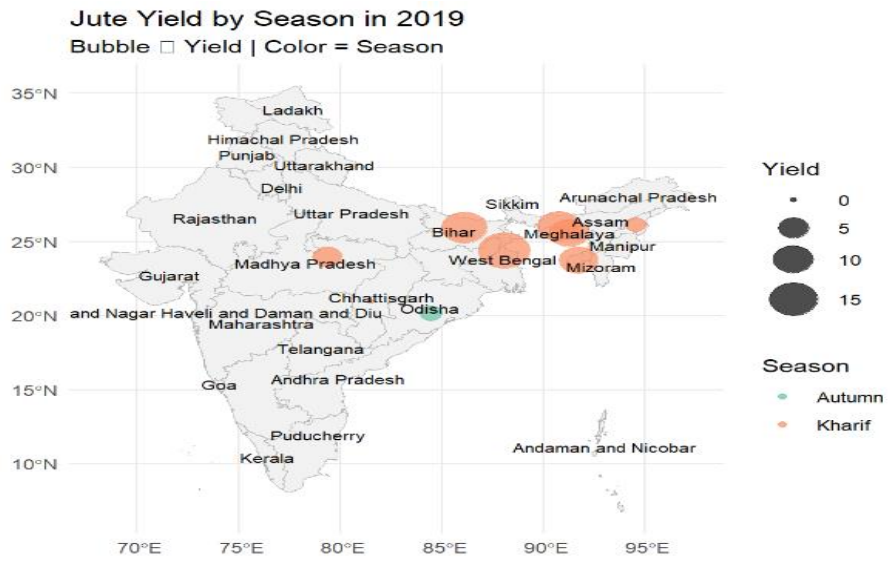
States like Kerala, Assam, and Arunachal Pradesh adopted "Kharif " sugarcane.

A steady increase in yield is visible from 1997 to 2019, particularly in states like Maharashtra, Andhra Pradesh, and Uttar Pradesh. Additionally, the adoption of "Whole Year" cultivation has expanded significantly over time.

2.1.4 Analysis of Jute







- **In 1997 :**

During the "**Kharif**" season, West Bengal, Assam, and Bihar primarily produced Jute.

- **In 2008 :**

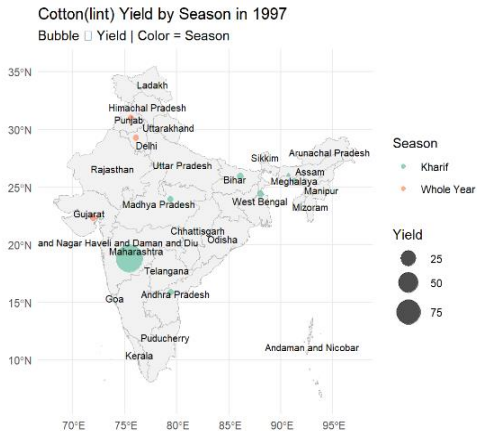
West Bengal remained dominant, but Odisha entered with jute cultivation in the Autumn season.

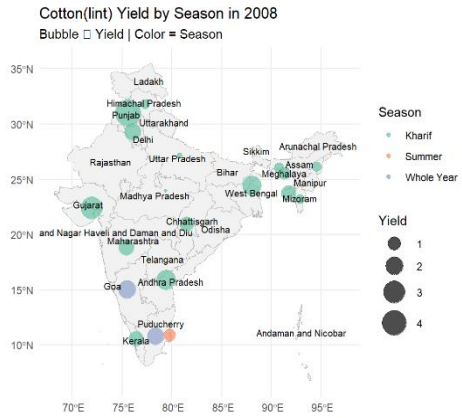
- **In 2019:**

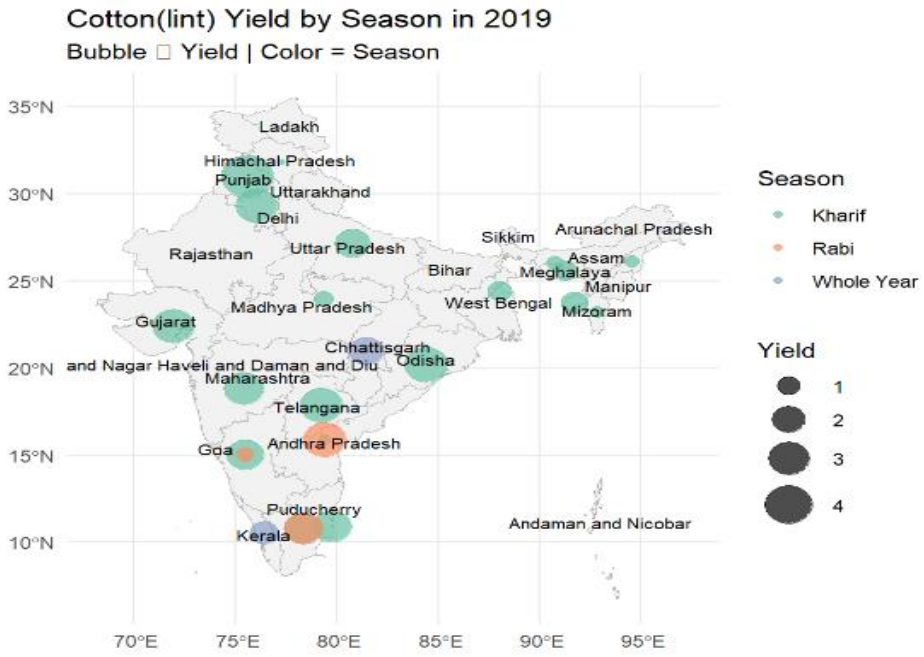
Production became more intense, particularly in West Bengal, with noticeable growth in Assam and Bihar. Most states continued with "Kharif".

West Bengal has consistently been the highest-yielding state, reflecting its agro-climatic suitability and longstanding jute industry.The increasing yield in states like West Bengal and Assam over the decades highlights regions of high productivity. Meanwhile, the rare presence of Autumn-season jute suggests limited seasonal variation in jute farming across India.

2.1.5 Analysis of Cotton







- **In 1997**

Only "Kharif" and "Whole Year" seasons are recorded.

States like Maharashtra and Telangana had higher yields in the "Kharif" season.

- **In 2008**

States like Odisha, Assam, and even Delhi are involved.

"Rabi" season cotton appears in Kerala.

- **In 2019**

Cotton farming has further diversified and reached more states like Himachal Pradesh, Bihar, West Bengal, etc.

Andhra Pradesh shows "Rabi" season cotton yield .

Geographic spread of cotton farming increased significantly from 1997 to 2019. Yield intensity was higher in core cotton-growing states like Maharashtra and Gujarat throughout the years. By 2019, cotton farming became more widespread across both northern and eastern states.

2.2 Spatial autocorrelation

The spatial autocorrelation concept is that it represents the relationship between nearby spatial units, as seen on maps, where each unit is coded with a realization of a single variable.

Moran's I (Global Spatial Autocorrelation)

Tests whether a variable is **clustered, dispersed, or random** across space.

Formula:

- $$I = \frac{n}{W} \cdot \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

where:

n: number of spatial units
x_i: value at location i
$\bar{x}$: mean of the variable
w_{ij}: spatial weight between units iii and j
W: sum of all w_{ij}
Interpretation:

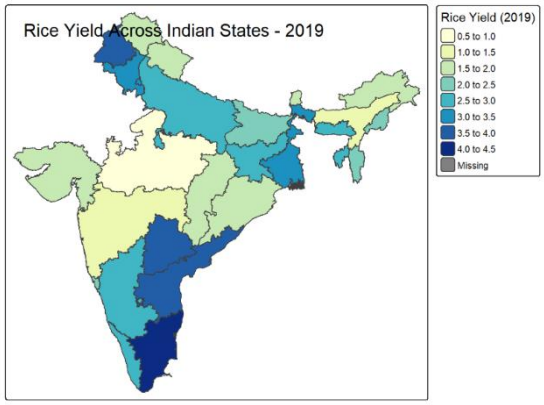
$I>0$: positive spatial autocorrelation (clusters)
$I<0$: negative spatial autocorrelation (dispersion)
$I \approx 0$: random pattern

LISA (Local Indicators of Spatial Association)

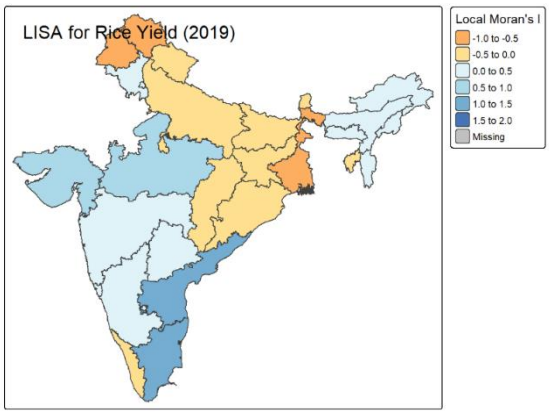
While **Moran's I** is global, **LISA** detects **local clusters or outliers** in spatial data.

Local Indicators of Spatial Association (LISAs) are statistical tools that identify areas where observed values differ significantly from the broader spatial pattern, revealing localized clusters (hot spots) or areas with dissimilar values (outliers). They are used to analyse spatial data, particularly in geographic information systems (GIS), and help in understanding the distribution of phenomena across space. The LISA for each observation gives an indication of the extent of significant spatial clustering of similar values around observation.

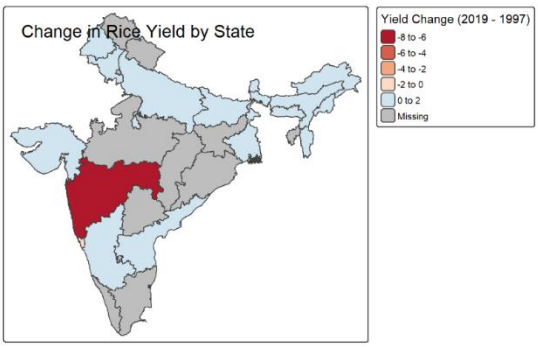
Rice



Southern states like Tamil Nadu and Andhra Pradesh achieved the highest rice yields, with values between 3.5 to 4.5. Northern and western states (e.g., Gujarat, Rajasthan, parts of Maharashtra) had lower yields, in the 1.0 to 2.0 range. Central and eastern belt (e.g., Odisha, Chhattisgarh) shows moderate productivity, around 2.5 to 3.0.



Southern states (Tamil Nadu, Karnataka, Andhra Pradesh) show positive autocorrelation → their high yields are matched by neighbouring states. Punjab, Haryana, parts of NE India show low or no correlation, possibly acting as spatial outliers or more independent in yield. Maharashtra and some central states show weak or negative spatial association, indicating they deviate from their neighbours.

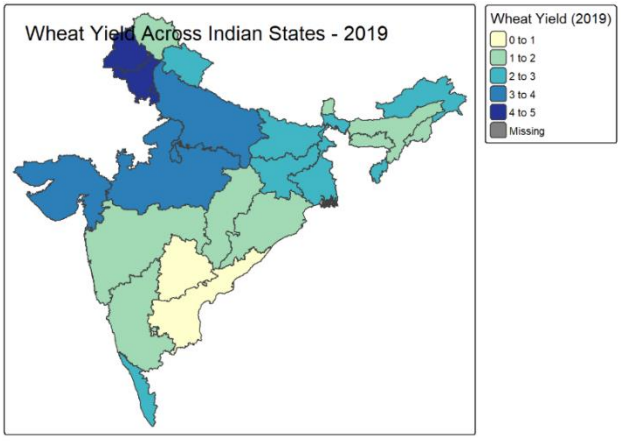


Maharashtra saw the sharpest decline in rice production (due to factors like monsoon patterns, water availability, and specific climatic conditions). States in the northeast and south show mild increases or stability.

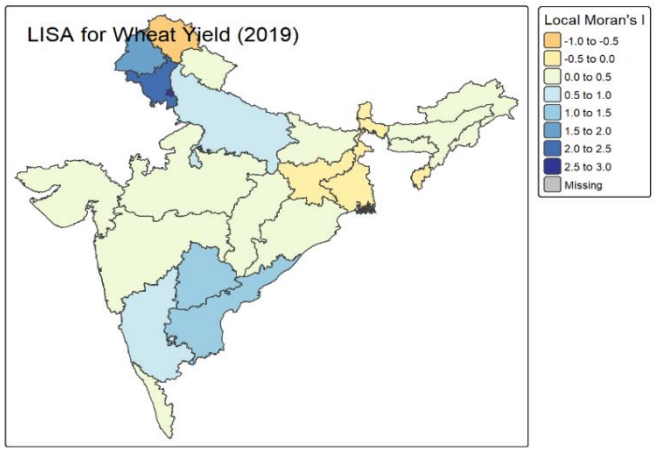
```
Moran I statistic standard deviate = 3.7833, p-value = 7.739e-05
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.202451124      -0.016949153       0.003363081
```

There is strong positive spatial correlation between Rice Production . Nearby regions tend to have even more strongly clustered rice production levels.

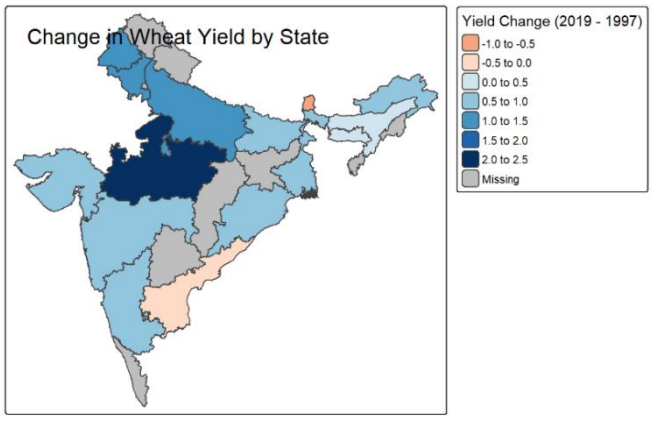
Wheat



Punjab has the highest wheat yield in the country Haryana, Uttar Pradesh, and Madhya Pradesh also grow a lot of wheat. Southern states like Karnataka, Tamil Nadu, and Maharashtra grow less wheat.



Punjab and Haryana have high wheat yield and their neighbours do too. They form a strong group. Some states like Jharkhand and Odisha (yellow/orange) have different yields compared to them. This map tells us where wheat success is clustered and where it's more random.



Madhya Pradesh improved the most , its wheat yield increased .Some southern states (like Tamil Nadu) saw a drop in wheat yield.

```
Moran I statistic standard deviate = 3.622, p-value = 0.0001462
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
       0.50061628       -0.04166667        0.02241554
```

There is moderate positive spatial autocorrelation in wheat production.

This means regions that are geographically close tend to have similar levels of wheat production.

The very small p-value indicates this pattern is statistically significant, so it's unlikely to be random.

This could reflect the influence of similar climate, soil, or farming practices in neighboring areas**.**

2.3 FORECASTING USING ARIMA

Moving from EDA we now come to predict or forecast the yield of some of the major crops for some of the different states of India. We start with the **Sweetest Part of India, West Bengal**.

2.3.1Forecasting For West Bengal:

According to the financial reports, Ministry of Agriculture, Govt. of India's various report it is easily understandable that the economy of Bengal is more or less driven by agricultural sector , by exporting the agricultural products to the different parts of India even foreign country. In this exporting of agricultural crops one of some major crops that boost the agricultural economy of Bengal is Rice. So it is very much important to forecast the production or yield for the basic policy making and future planning related to agriculture for the state.

Yield Forecast of Rice for West Bengal for 2020 to 2029:

We have data for the state of West Bengal, with yield values from the year 1997 to 2019.

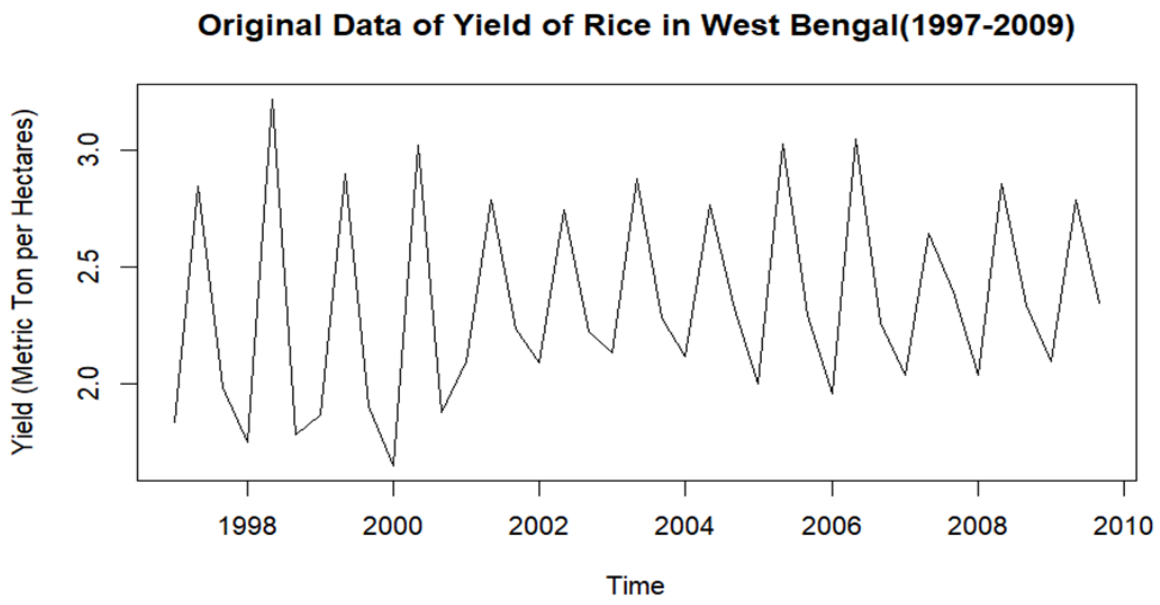
```
unique(wb_rice$Season)
```

```
[1] "Autumn     " "Summer     " "Winter     "
```

So we have 3 unique seasons in each year for production of rice which are autumn, summer, and winter. Using this dataset, we aim to forecast the yield values for the years 2020 to 2029.

We aim to forecast the data, so our main objective is to fit a model that minimizes the Mean Squared Error (The MSE on the test data evaluates the model's generalization ability by measuring the average squared difference between the predicted and actual values in the test set) of the forecast. To search for such a model, we first divide the available data into two parts: 1997-2009 and 2010-2019. We fit a model using the first part of the data, called training part, then generate predictions and compare them with the actual values from the second part to evaluate the MSE. For this purpose, we convert both parts of the data into time series with a frequency of 3, corresponding to the three seasons (Summer, Autumn and Winter) per year. Here, we use data from the years 1997 to 2009 as the training set, and data from 2010 to 2019 as the test set.

Here we have the time series plot for yield of rice in West Bengal from 1997 to 2009 .(in this case this is our training data.)



Examine the stationarity of the time series data:

To test the stationarity of our time series data, we use the Augmented Dickey-Fuller test.The ADF test specifically tests for the presence of a unit root in the series. If a unit root is present, the series is non-stationary . If the test rejects the null hypothesis, it suggests the series is weakly stationary.

Augmented Dickey - Fuller Test

The Augmented Dickey-Fuller (ADF) test is a statistical test used to determine whether a time series data set is stationary or non-stationary. The ADF test helps determine whether differencing is necessary to make the time series stationary. The ADF test evaluates the null hypothesis that a unit root is present against the alternative hypothesis of stationarity. If the test statistic obtained from the ADF test is smaller than the critical value at a given significance level, the null hypothesis of a unit root is rejected, suggesting the presence of stationarity in the data.

The ADF test is applied to the model

$$\Delta X_t = \alpha + \beta t + \gamma X_{t-1} + \delta_1 \Delta X_{t-1} + \cdots + \delta_{p-1} X_{t-p+1} + \epsilon_t$$

where α is a constant, β the coefficient on a time trend and p the lag order of the autoregressive process. Imposing the constraints α=0 and β=0 corresponds to modelling a random walk and using the constraint β=0 corresponds to modelling a random walk with a drift.

The unit root test is then carried out under the null hypothesis γ=0 against the alternative hypothesis of γ<0 .

Once a value for the test statistic: $DF_\tau = \frac{\hat{\gamma}}{SE(\hat{\gamma})}$

Where, $\hat{\gamma}$ is the estimated value of γ . As this test is asymmetrical, we are only concerned with negative values of our test statistic DF_τ. If the calculated test statistic is less (more negative) than the critical value, then the null hypothesis of γ=0 is rejected and no unit root is present.

```
adf.test(wb_rice_ts_1)
```

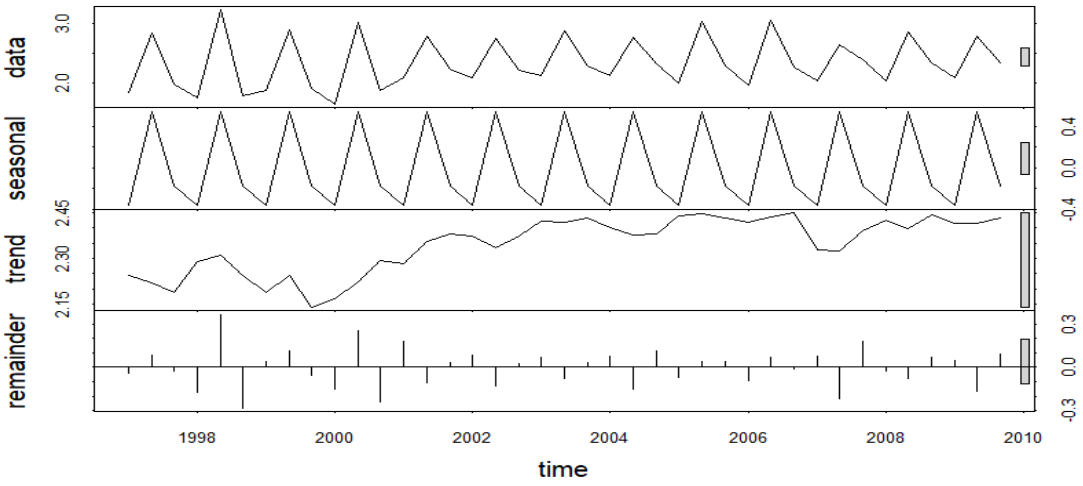
```
        Augmented Dickey-Fuller Test

data:  wb_rice_ts_1
Dickey-Fuller = -1.7269, Lag order = 3, p-value = 0.6803
alternative hypothesis: stationary
```

Since the p-values for all datasets exceed the 0.05 significance level, we do not have sufficient evidence to reject the null hypothesis of a unit root. This indicates that, at the 5% significance level, all the datasets exhibit non-stationary behavior.

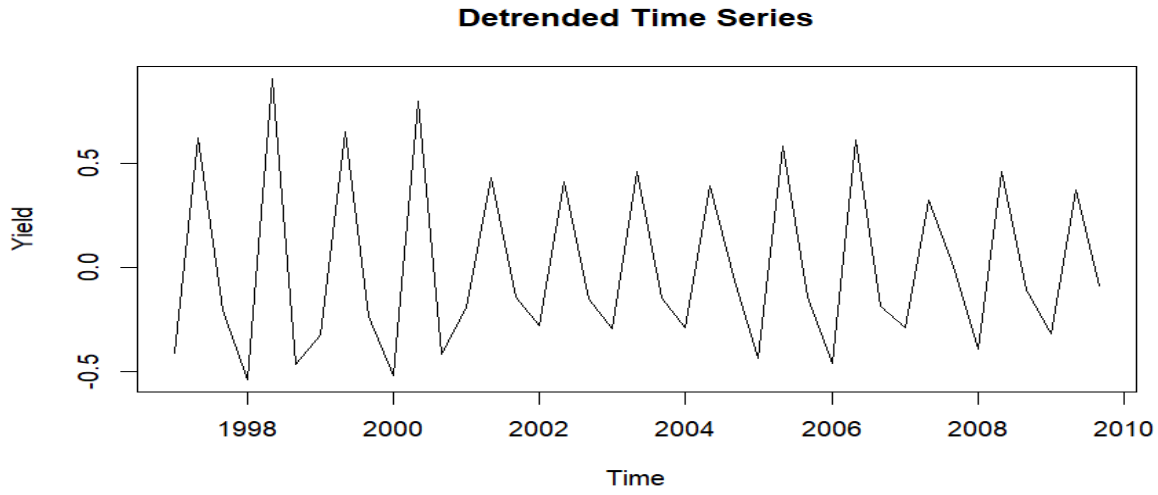
Component of time series:

- Trend
- Seasonal component
- Cyclical component
- Irregular (random noise) component.

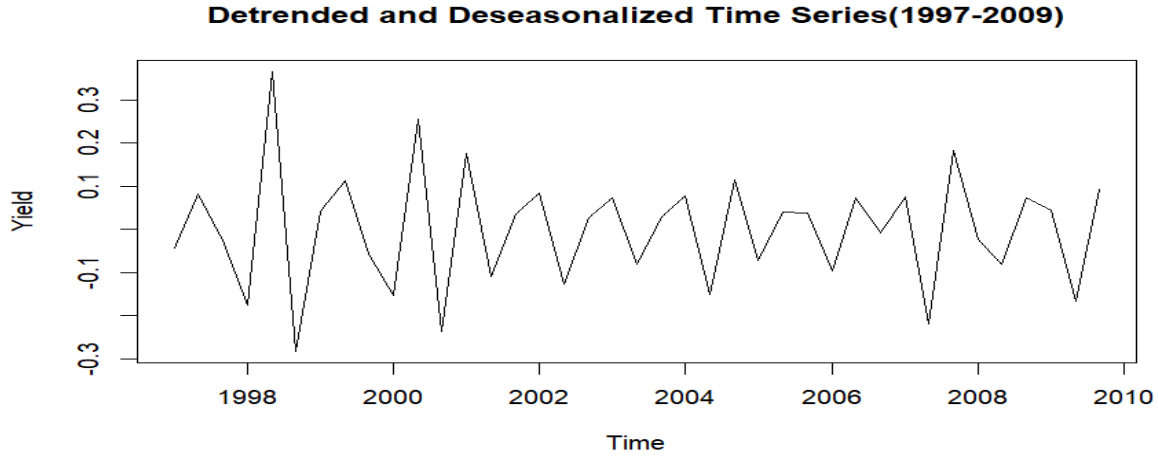
Our target is to find the staionarity in the dataset as the Autoregressive (AR) model concerning about Time Series to be Stationary. To do so we first check whether there is any trend and seasonlity present in the training data or not.



The graph shows that the decomposed time series data where it is clearly visible that trend and seasonality is present.
We now remove trend and seasonality from the time series data.



Our next task is to remove seasonality from this data.



Now we have our detrended and deseasonalised time series data for Yield of rice in West Bengal from the year 1997 to 2009.

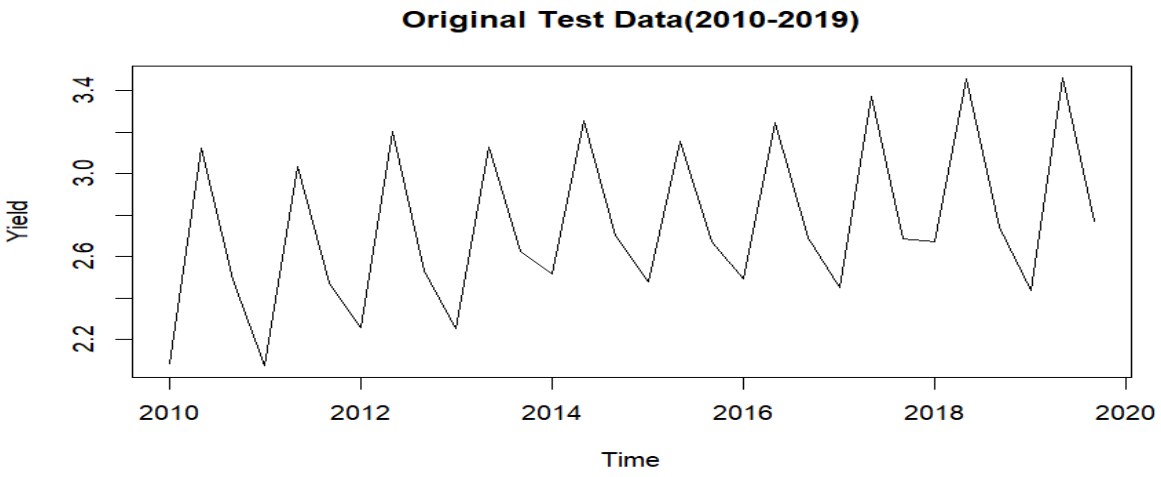
The ADF Test for checking the stationarity of this detrended and deseasonalised time series data are given below:

```
adf.test(wb_rice_ts_new_1)
```

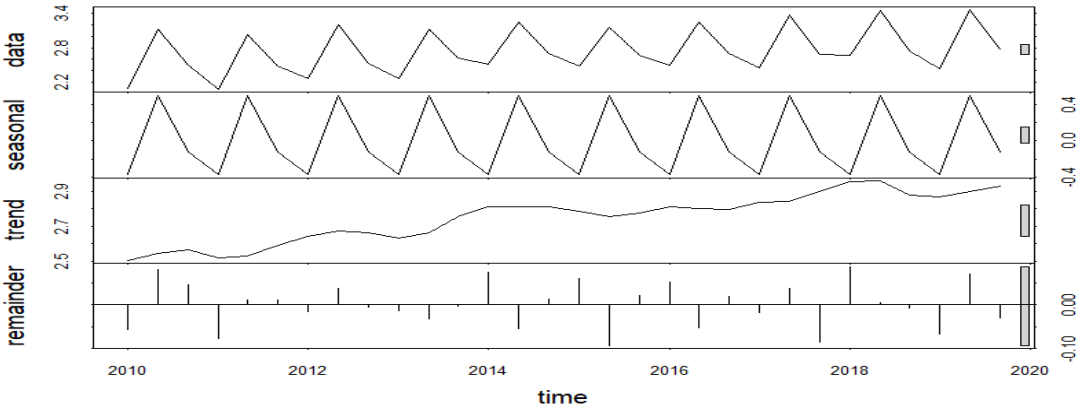
```
        Augmented Dickey-Fuller Test

data:  wb_rice_ts_new_1
Dickey-Fuller = -6.5383, Lag order = 3, p-value = 0.01
alternative hypothesis: stationary
```

As the p-value of the test is 0.01 which is clearly <0.05 this shows we can reject our null hypothesis of a unit root at 5% level of significance. We can now say that our time series training data is stationary.

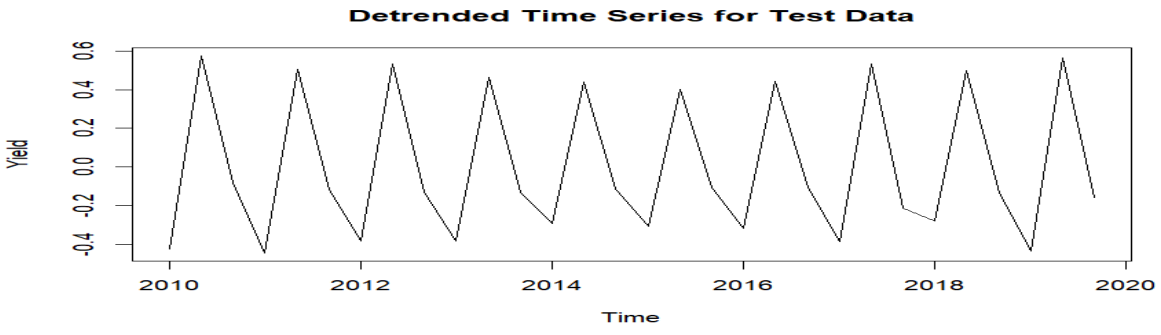
Repeating the same procedure for our test set which is yield of rice in West Bengal from 2010 to 2019;



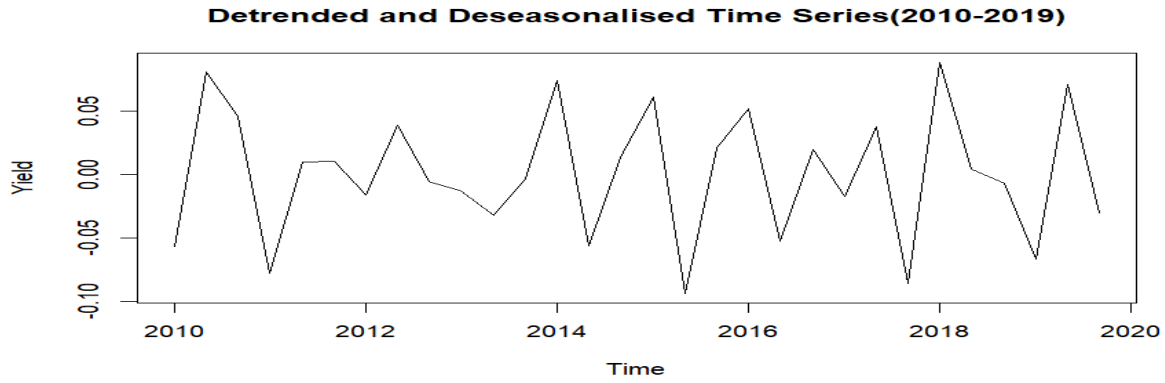
Decomposing the time series we get the following:



Again we remove trend for the test data.



The above plot shows us the detrended data where seasonality is present . We should go for removing seasonality.



Finally we get our detrended and deseasonalised time series data for the test set which is yield of rice in West Bengal from the year 2010 to 2019.

Autoregressive(AR) Model:

An autoregressive process of order p, denoted AR(p), models the current value of a time series as a linear combination of its p previous values plus a random error. If Z_t is a white noise process with mean zero and variance ${\sigma_z}^2$, then X_t is said to follow an AR(p) process if:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + Z_t$$

where $\theta_{i'}s$ are constants. For the AR(p) process to be valid, must be weakly stationary X_t.

Moving Average(MA)Model:

A moving average process of order q, denoted MA(q), models the current value of a time series as a linear combination of the current and past q white noise error terms. If Z_t is a white noise process with mean zero and variance ${\sigma_z}^2$, then X_t is said to follow an MA(q) process if:

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q'}$$

where $\theta its(i = 1, 2, \cdots, q)$ are constants.

Autoregressive Moving Average (ARMA(p,q)) Model:

The process $X_t; t = 0, \pm 1, \pm 2, \ldots$ is said to be an ARMA(p, q) process if X_t is stationary and if for every t,

$$X_t - \cdots \phi_1 X_{t-1} - \ldots . - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + + \theta_q Z_{t-q'}$$

where $Z_t \sim WN(0, \sigma^2)$.

Autoregressive Integrated Moving Average (ARIMA) Model:

If X_t is the original time series, the ARIMA(p, d, q) model is written as:

$$\left(1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p\right)(1 - B)^d X_t = (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q)\varepsilon_t$$

Where :

B is the backshift operator

The equation fitted is a simple linear trend model, i.e.,

$$X_t = \beta_0 + \beta_1 t + \varepsilon_t$$

Where,

X_t is the observed value at time t,

β_0 is the intercept,

β_1 is the slope,

ϵ_t is the random error

Finally we are ready to fit ARIMA model with our staionary training dataset (Yield of Rice in West Bengal from 1997 to 2010)

Our objective is to get an appropriate model, to do so we see the MSE's for ARIMA model of different orders and select the model with least MSE.

Here we assume the parameters of ARIMA (p,d,q) model in a given range of values and latter calculate the MSE for each of them. We take values p=8, d=3 and q=5

After model fitting, to evaluate forecasting accuracy on test data we see the order of the model for which minimum MSE is observed. Here the minimum MSE is observed for the model with order (p,d,q)≡(2,1,1)

We now fit the ARIMA model of the training dataset with order (2,1,1)

```
summary(arima_model)
```

```
Call:
arima(x = wb_rice_ts_new_1, order = c(2, 1, 1))

Coefficients:
          ar1      ar2      ma1
      -1.1881  -0.7142  -1.0000
s.e.   0.1065   0.1052   0.0734

sigma^2 estimated as 0.004268:  log likelihood = 45.84,  aic = -83.68

Training set error measures:
                      ME       RMSE        MAE       MPE     MAPE      MASE        ACF1
Training set -0.003393114 0.06448771 0.04998654 7.729991 62.55312 0.2487948 -0.4435564
```

As we can see the smaller standard error values and smaller variance of residuals (sigma^2) indicate the better fit of the model.

Akaike Information Criterion

AIC are widely used in model selection criteria. AIC means Akaike Information Criteria. The AIC can be termed as a measure of the goodness of fit of any estimated statistical model. The model with the lowest

AIC offers the best fit.

$$AIC = -2 * log(L) + 2 * k$$

In the formula, L represents the maximized likelihood of the model, which measures how well the model fits the data. The term k represents the number of parameters in the model, including the intercept and any additional predictors. the negative log-likelihood is Inf.

In the given model the lower value of AIC suggest the better fit of the ARIMA model with order (p,d,q)≡(2,1,1)

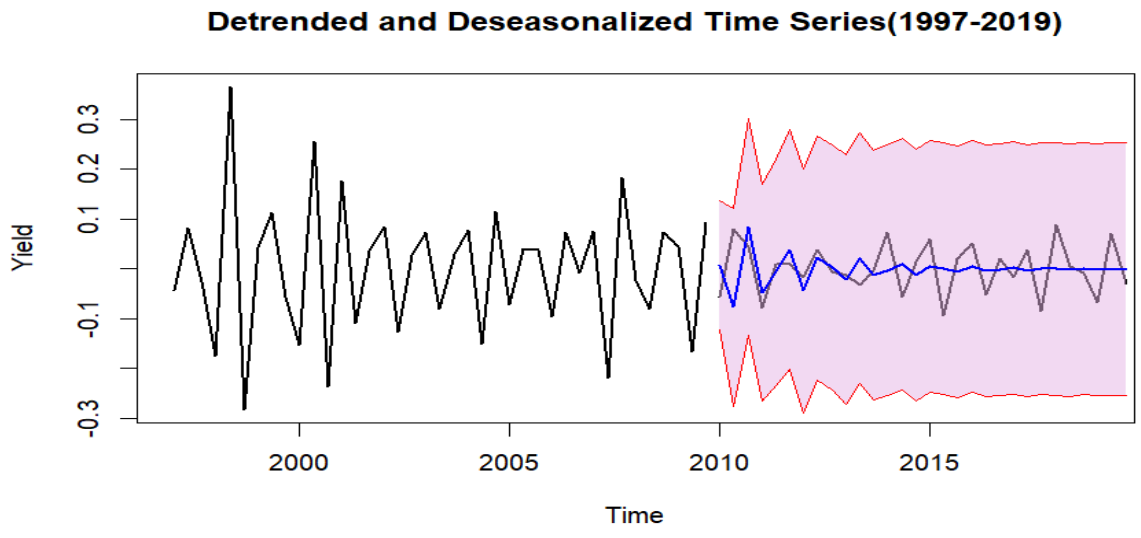
Now we compare these forecasted values with the actual values that we have in terms of our test data set i.e. the time series data for the Yield of rice in Bengal from the year 2010 to 2019 and we see the difference between them.

The average of the squared of these errors of forecasting values i.e. the MSE is:

```
cat("Model MSE",mean((wb_rice_ts_new_2-forecast_val$mean)^2))
```

```
Model MSE 0.003103114
```

We had our detrended and deseasonalised and stationary time series data for the test dataset which is the yield of rice in Bengal from the year 2010 to 2019. Now comapring this test set with our predicted values we get the following:



In this graph it is shown that the actual values of the yield from 2010 to 2019 in black colour and the predicted / forecasted values of yield from 2010 to 2019 in blue colour and it is also observed that our actual yield values are lying in the forecasted confidence interval for the data by which we can say that our forecasting is more or less reliable.

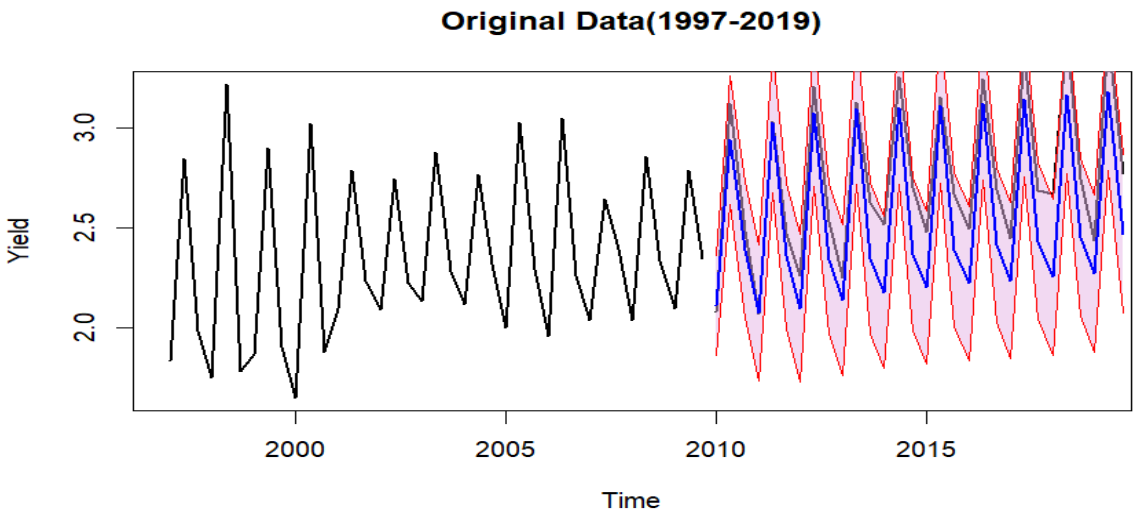
Now inlcuding trend and seasonality in the dataset we comapre the actual and forecasted values.

Finally we get the original values of time series test data (with trend and seasonality included) , forecasted values and forecasted upper and lower limit of the 95% confidence interval.

The average of the squared of these errors of forecasting values i.e. the MSE is:

0.05021837

The lower MSE suggests the better forecasting.



The curve shows us the original data and predicted data of yield of rice in west bengal from 2010 to 2019 which has both trend and seasonality

Similar to the previous plot the black line suggests the original yield values (including trend and seasonality) , blue line suggests the forecasting values of yield of the test data with trend and and seasonality included in the data the two red lines suggests the upper and lower limits of the forecasting values of the test dataset from 2010 to 2019 which includes trend and seasonality.

Keeping in mind this whole procedure we now fit the ARIMA model for whole dataset i.e. the yield of rice in West Bengal from the year 1997 to 2019 and forecast the future values for upcoming 10 years including 3 seasons each.

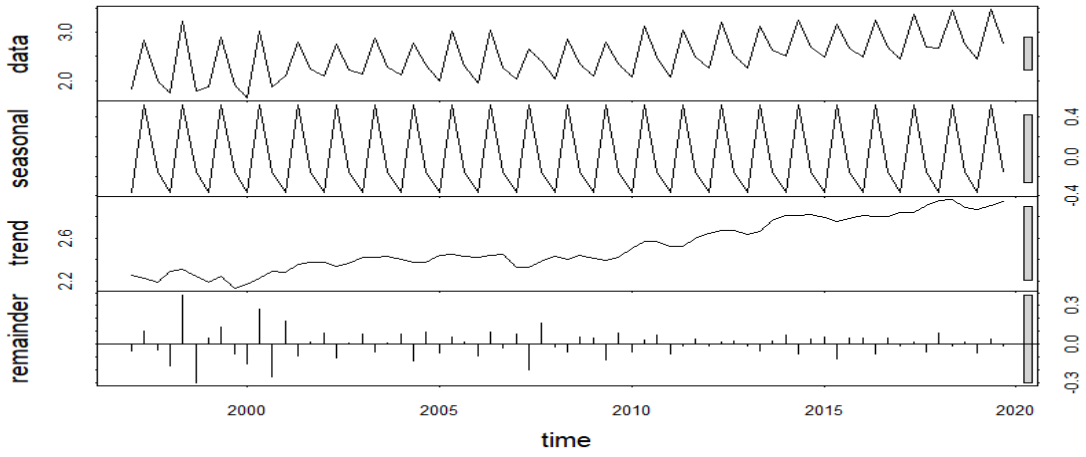
This is the plot for the original time series data we have in our hands .

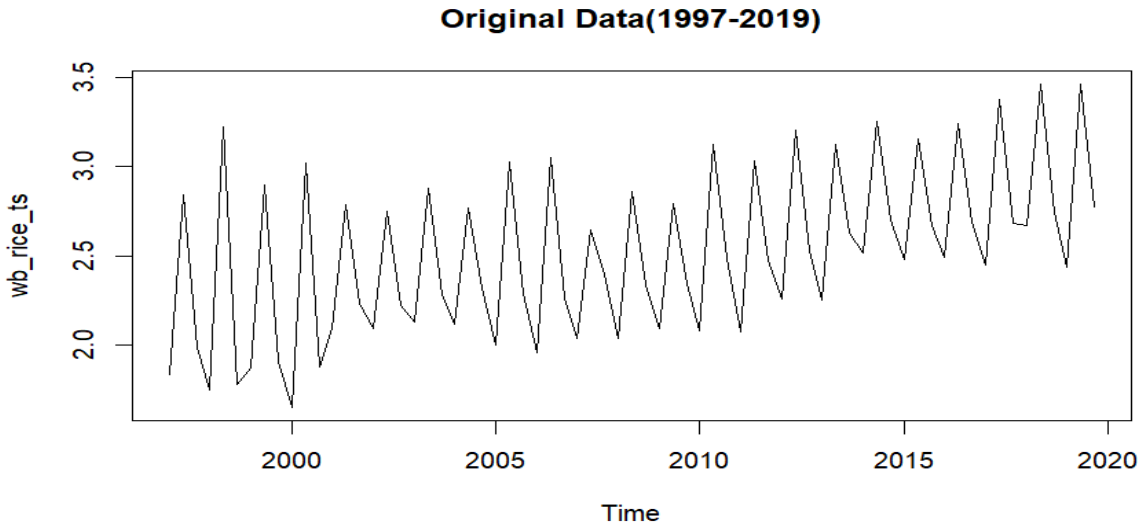
We see that the time series is stationary.

```
adf.test(wb_rice_ts)
```

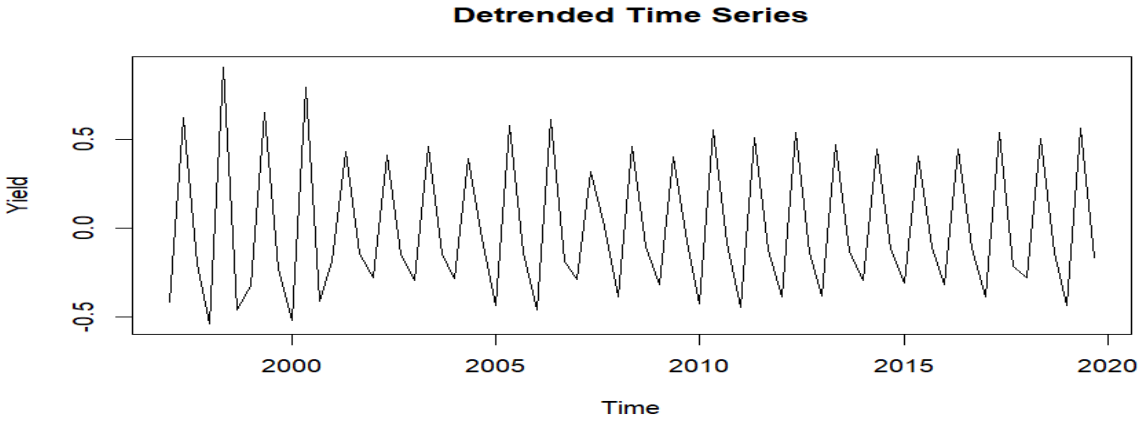
```
        Augmented Dickey-Fuller Test

data:  wb_rice_ts
Dickey-Fuller = -3.2366, Lag order = 4, p-value = 0.08923
alternative hypothesis: stationary
```

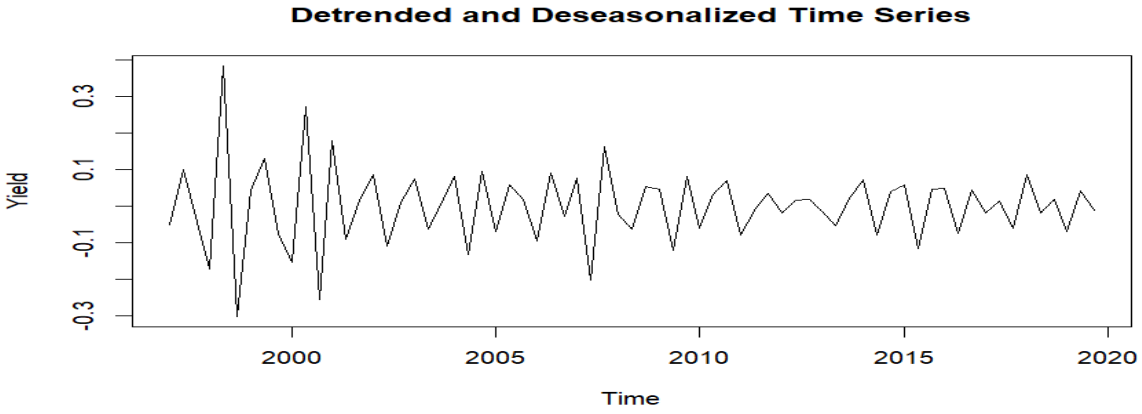
Decomposing the data;





we remove the trend and seasonality;





Now we test for stationarity of this detrended and de seasonalised time series data:

```
adf.test(wb_rice_ts_new)
```

```
        Augmented Dickey-Fuller Test

data:  wb_rice_ts_new
Dickey-Fuller = -8.1373, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

As the p-value of the test is 0.01 which is clearly <0.05 this shows we can reject our null hypothesis of a unit root at 5% level of significance. We can now say that our time series data is stationary.

Now we fit the ARIMA model of order (2,1,1) ,which we got from previous analysis, for the whole dataset from1997 to 2019 .

Model AIC =

-192.89

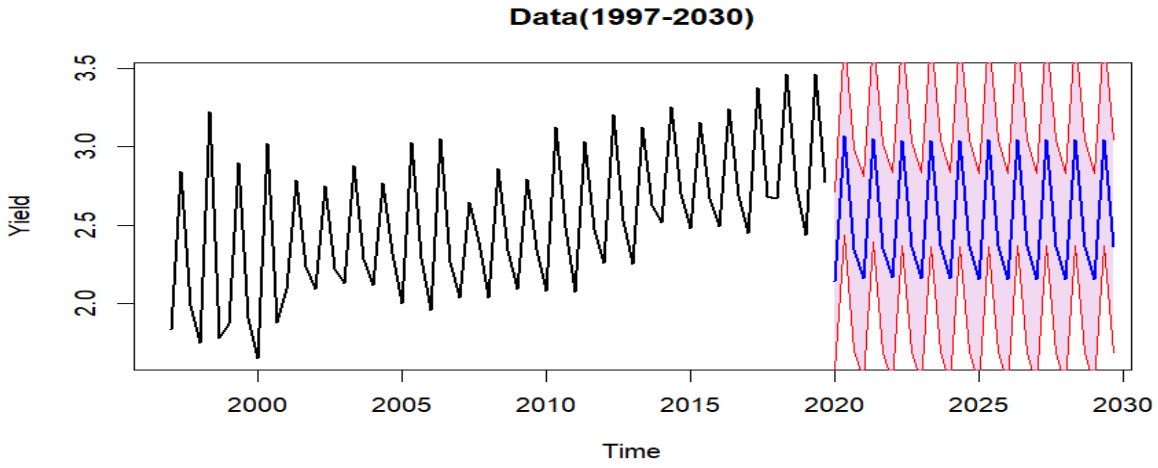
Estimated model Sigma^2 =

sigma^2 estimated as 0.002945

The lower value of sigma^2 and AIC suggests that our model fits the data well.

Again we include the trend and seasonlity in the original and predicted data .

So we have the data for the original values of yield from 1997 to 2019 ; predicted values of yield from 2020 to 2030 ; the upper and lower 95% confidence interval for the predicted yield values.



Finally we have the desired plot that includes the time series data for the original dataset of yield from 1997 to 2019 ; predicted values of yield for 10 years i.e. from 2020 to 2029.

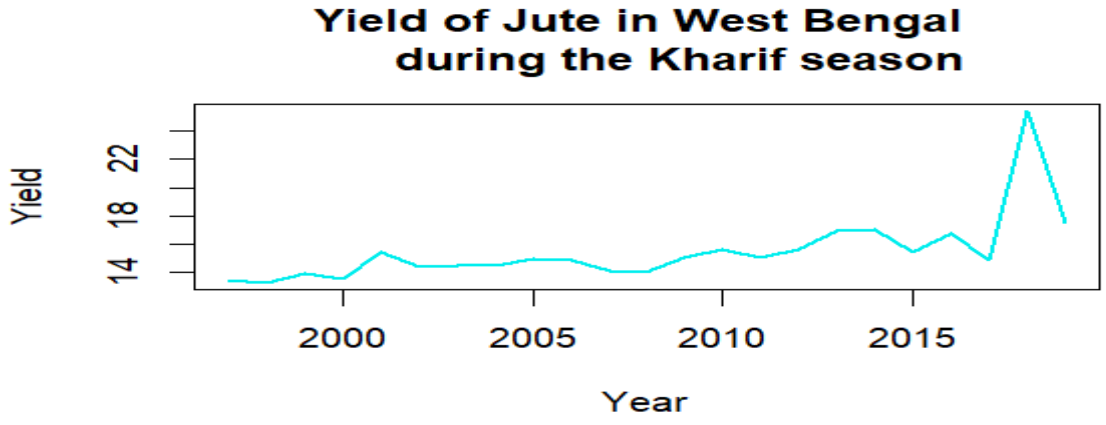
The black line suggests the original dataset from 1997 to 2019 and blue line suggests the forecasting values from 2020 to 2029; the red lines suggest the upper and lower 95% confidence interval for the predicted values and we can say that our predicted values for the yield from 2020 to 2029 are perfectly allign in the confidence interval, which suggests that our model is reliable.

Yield Forecast of Jute for West Bengal for 2020to 2029:

West Bengal is known for its jute production, particularly along the border with Bangladesh and south of the Ganges River. It is the 2nd largest crop in West Bengal in terms of Production rate or yield. After rice it is the second most important exporting product that boost up the State's aggricultural economy.

We have the data of Yield of Jute from 1997 to 2019. As it is cultivated only in one season i.e. Khariff so we have the data only for one season in this case.

The time series data of Yield of Jute from 1997 to 2019 is given below:



For forecasting the Jute yield for upcoming 10 years in West Bengal we proceed in a similar way as we done in the previous section we processed the data in a similar way we first divide the time series data into two part training data which is the Yield of jute in West Bengal from 1997 to 2009 and second is test data set which is the yield of Jute from 2010 t 2019. Then we check for seasonlity and trend and checking whether it is a stationary time series or not. After these methods when we have detrended deseasonalised staionary data in our hand ; we are going for ARIMA fitting for the training dataset as we have done before.

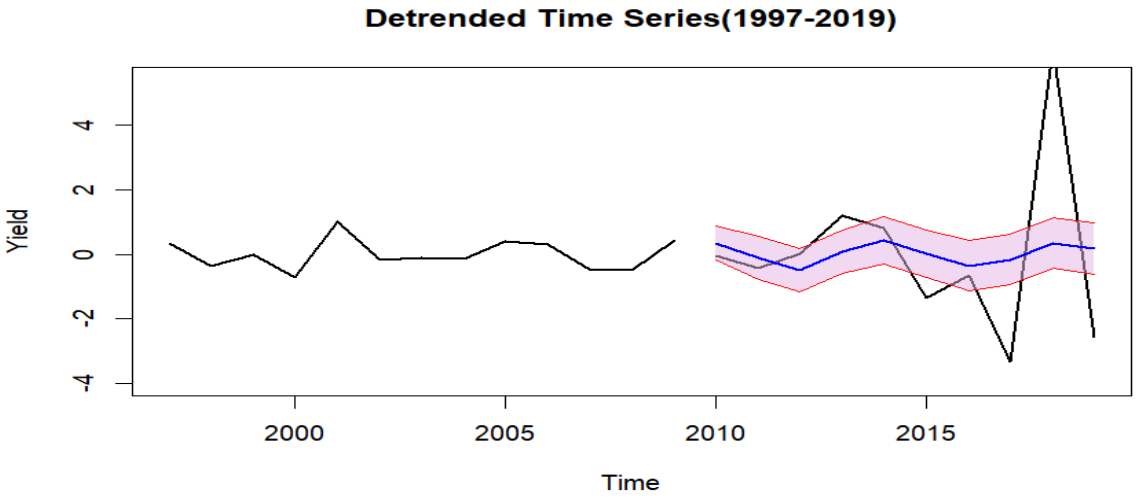
Once again our objective is to find the appropriate model and for that we use MSE to find the order of the best model fitted by ARIMA. Using the parameter for which fitted arima model gives the smallest mse , we forecast the values of yield for the test set i.e. the yield values for Jute from the year 2010 to 2019.

```
forecast_val=forecast(arima_model,h=10)
```

```
Call:
arima(x = wb_jute_ts_new_1, order = c(5, 0, 0))

Coefficients:
          ar1      ar2      ar3      ar4     ar5  intercept
      -0.7884  -0.5926  -0.5841  -0.0228  0.3411    -0.0015
s.e.   0.2845   0.4239   0.4520   0.4938  0.3559     0.0345

sigma^2 estimated as 0.07133:  log likelihood = -2.92,  aic = 19.83

Training set error measures:
                      ME      RMSE       MAE       MPE     MAPE      MASE        ACF1
Training set -0.002098744 0.2670682 0.1949225 18.68256 64.86747 0.3312888 0.08508475
```

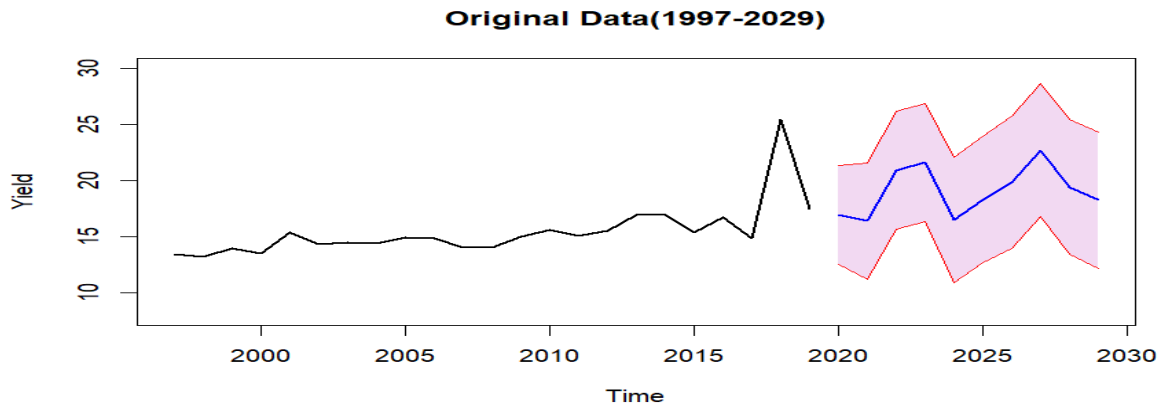
This value of AIC suggest more or less a good fitting of the training data.

Again keeping in mind the previously used method we have our forecasted value for time interval 2010 to 2019 and have also the original value of yield in this time period. We now find MSE (Mean Squared Error) of this model and then plot it .



This curve shows that our fitting for detrended deseasonalised test data is not so well.

We include the trend and seasonality compositions into the time series and finally we fit the whole data in ARIMA and predict the future values.



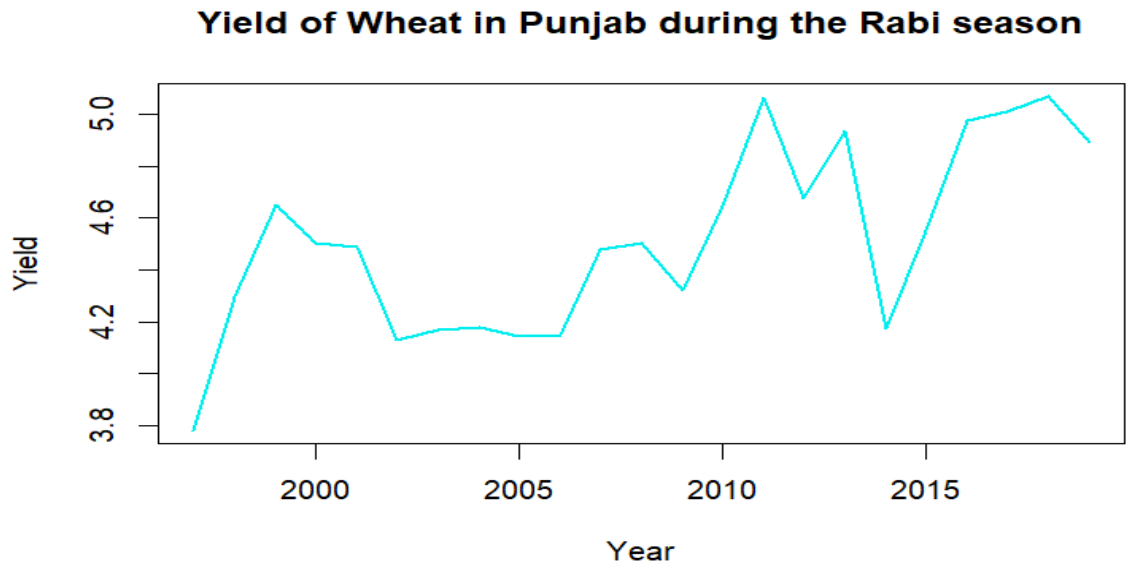
So we have the forecasting for yield of Jute in west Bengal for future time.

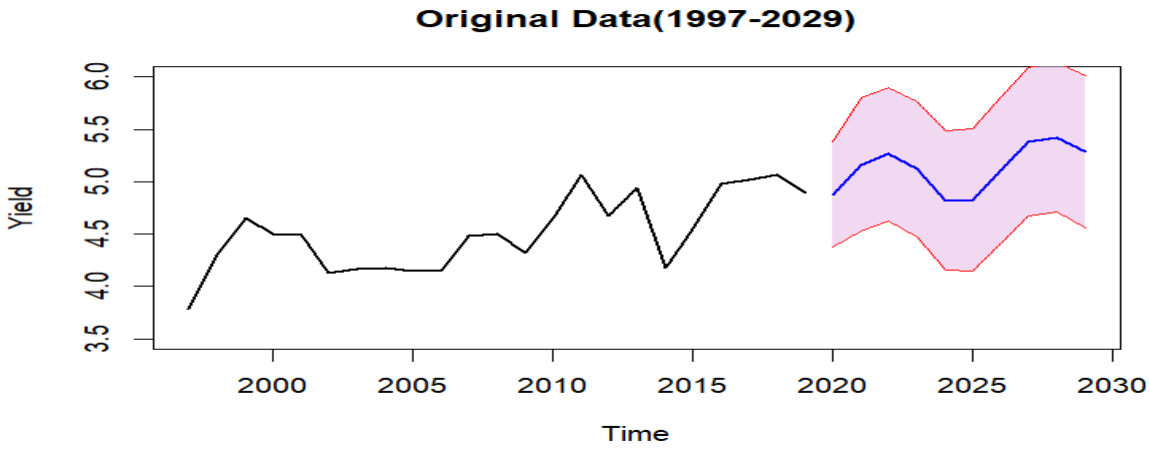
Black line indicates the original data (whole data i.e. yield from 1997 to 2019) blue line indicates the predicted data for the yield of jute from 2020 to 2029. Two red lines denotes the 95% confidence interval for the forecasted values. As we see that our predicted values allign in the interval which suggest that more or less a good prediction of our model.

Punjab:

Wheat

Plot for time series data for the Yield of Wheat in Punjab is given below:



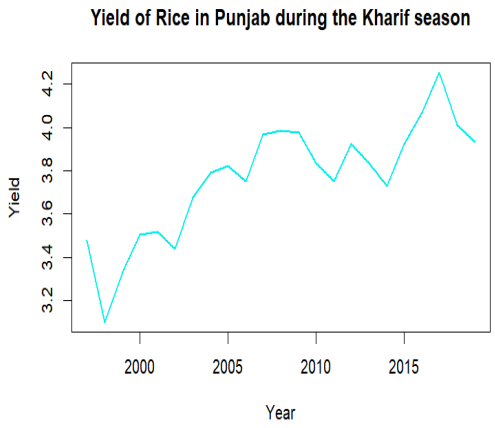


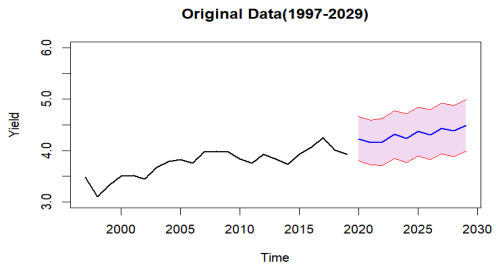
This grapgh shows that forecasted values of Yield for Wheat in Punjab during the rabi season.

Black line indicates the original data (whole data i.e. yield from 1997 to 2019) blue line indicates the predicted data for the yield of wheat from 2020 to 2029. Two red lines denotes the 95% confidence interval for the forecasted values. As we see that our predicted values perfectly allign in the interval which suggest that a good prediction of our model for the Wheat yield prediction in Punjab.

Rice

Plot for time series data for the Yield of Rice in Punjab is given below:





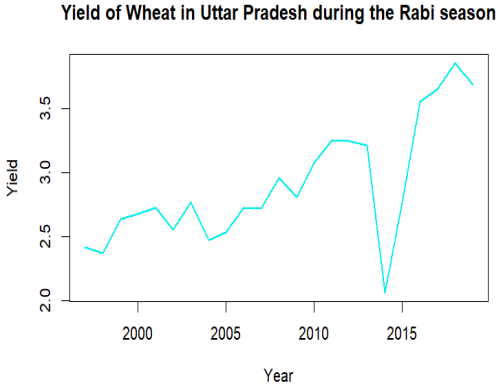
This grapgh shows that forecasted values of Yield for Rice in Punjab during the Kharif season.

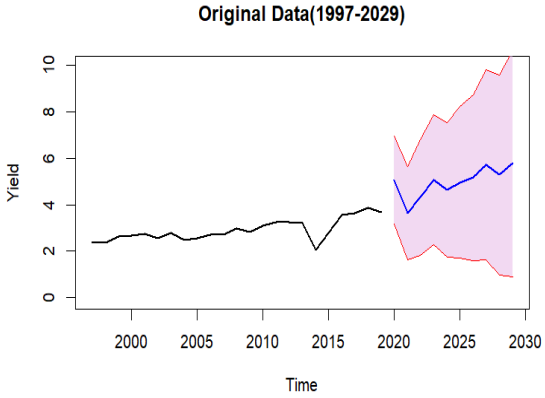
Black line indicates the original data (whole data i.e. yield from 1997 to 2019) blue line indicates the predicted data for the yield of Rice from 2020 to 2029. Two red lines denotes the 95% confidence interval for the forecasted values. As we see that our predicted values perfectly allign in the interval which suggest that a good prediction of our model for the Rice yield prediction in Punjab.

Uttar Pradesh

Wheat

Plot for time series data for the Yield of Wheat in Uttar Pradesh is given below:



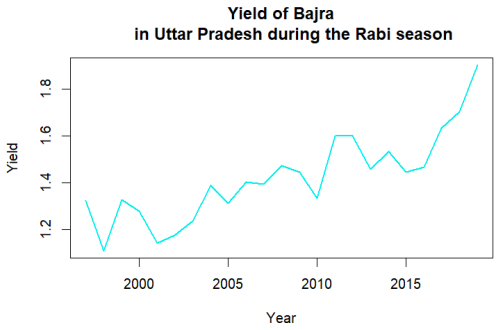


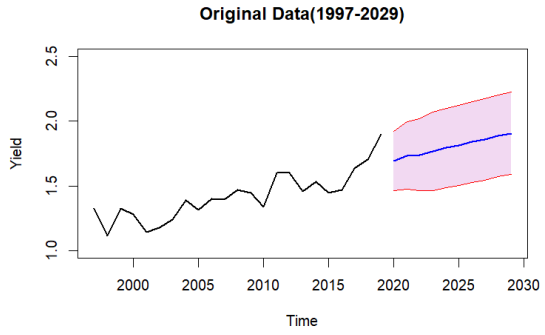
This grapgh shows that forecasted values of Yield for Wheat in Uttar Pradesh during the Rabi season.

As we see that our predicted values perfectly allign in the interval which suggest that a good prediction of our model for the Wheat yield prediction in Uttar Pradesh.

Bajra

Plot for time series data for the Yield of Bajra in Uttar Pradesh is given below:



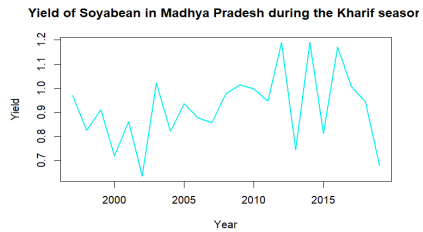


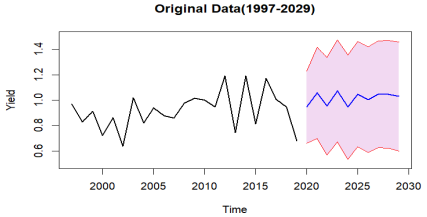
As we see that our predicted values for the future i.e. from year 2021 to 2029 are alligned in the interval which suggest that a more or less a good prediction of our model for the Bajra yield prediction in Uttar Pradesh.

Madhya Pradesh

Soyabeen:

Plot for time series data for the Yield of Soyabeen in Madhya Pradesh is given below:



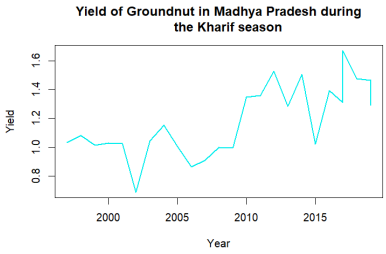


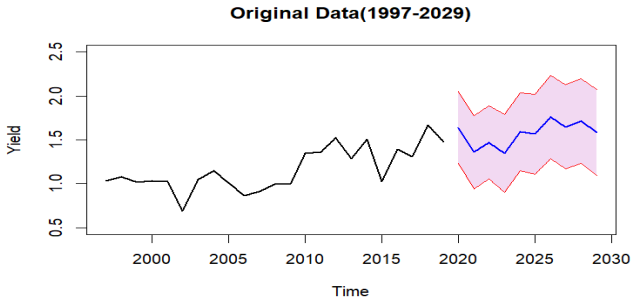
This grapgh shows that forecasted values of Yield for Soyabeen in Madhya Pradesh during the Kharif season.

Black line indicates the original data (whole data i.e. yield from 1997 to 2019) blue line indicates the predicted data for the yield of soyabeen from 2020 to 2029. Two red lines denotes the 95% confidence interval for the forecasted values. As we see that our predicted values are allign in the interval which suggest that a more or less good prediction of our model for the soyabeen yield prediction in Madhya Pradesh.

Groundnut:

Plot for time series data for the Yield of Soyabeen in Madhya Pradesh is given below:





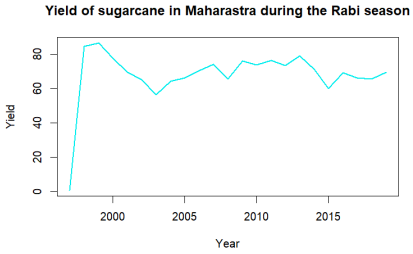
So we have the forecasting for yield of Ground Nut in Madhya Pradesh for future time.

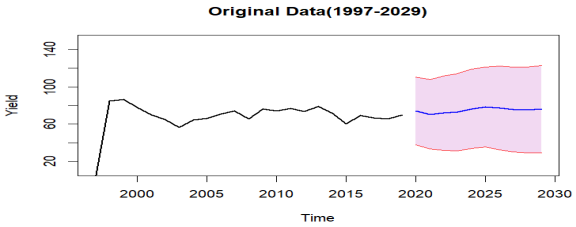
Black line indicates the original data (whole data i.e. yield from 1997 to 2019) blue line indicates the predicted data for the yield of groundnut from 2020 to 2029. Two red lines denotes the 95% confidence interval for the forecasted values. As we see that our predicted values perfectly allign in the interval which suggest that good prediction of our model.

Maharashtra

Sugarcane:

Like all other previous crops the time series data of yield of sugarcane in Maharashtra is plotted here:

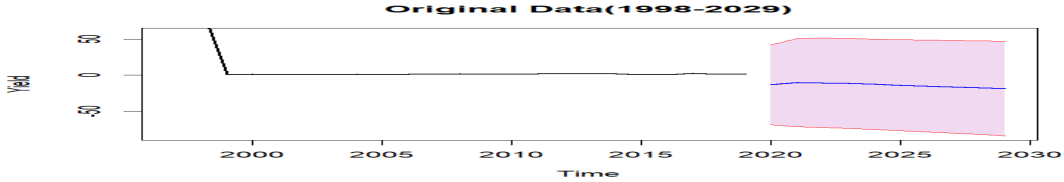




So we have the forecasting for yield of Sugarcane in Maharashtra for future time.

As defined earlier we can see that there is a good prediction reached by ARIMA model.

Cotton:



According to the curve and the previously gained knowledge from ARIMA model fitting we can say that there is a more or less good prediction shown in the yied value of Cotton in Maharashtra.

2.4 FORECASTING USING ARIMAX

2.4.1 ARIMAX MODEL

An ARIMAX model, which stands for AutoRegressive Integrated Moving Average with eXogenous inputs, is an advanced version of the ARIMA (AutoRegressive Integrated Moving Average) model. The ARIMAX model extends the ARIMA framework by integrating exogenous variables, which are external factors that can influence the time series being studied. This integration allows the model to leverage additional information that can significantly enhance forecasting accuracy.

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \beta x_t + \epsilon_t$$

where (Yt) is the value of the dependent variable at time (t,) (Φi)are the parameters of the autoregressive part, (θj) are the parameters of the moving average part, x_t represents the exogenous variables at time t, and $\boldsymbol{\beta}$ are the coefficients associated with these exogenous variables.

In the previous section we have discussed about the model fitting using ARIMA and got that the best fitted model and forecasted the values of yield of some crops. But we couldn't sure about the influence of the other exogeneous variable like Annual rainfall, pesticide, fertilizer etc. Now to analyse the effect of the covariates of the future predicted values of yield we shall now discuss the analysis for ARIMAX model.

2.4.2 ARIMAX For Yield of Rice in West Bengal:

Just like we have done before, first we analyse the effect of exogeneous variable on the Yield of Rice in West Bengal.

As we have done before for the basic ARIMA, here also we construct the time series data and split them into test and training set for each of the following variables: Annual Rainfall, Pesticide and Fertilizer.

As mentioned before the training data set include the Yield of rice from 1997 to 2009 with each exogeneous variable and test set includes the yield of rice in West Bengal from 2010 to 2019 with each of the exogeneous variables.

In case of each variable :

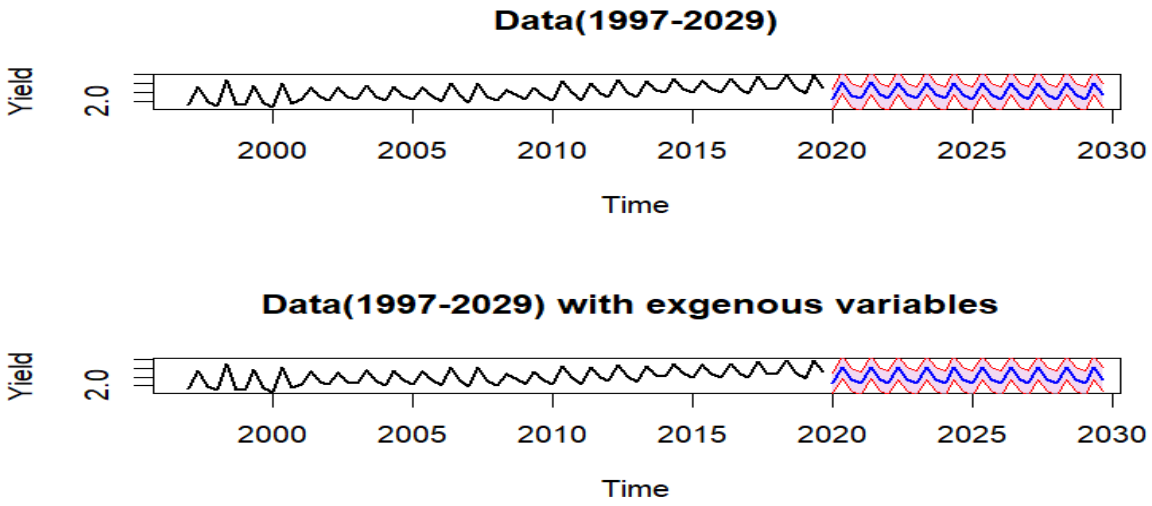
We go for stationarity check by adf test. If necessary then we decompose them by STL Decomposition and exclude the trend and seasonlity from the training dataset. Now we have the staionary data set with each exogenous variable. We now fit the ARIMAX model with variable Annual Rainfall for the training data set. Order of the ARIMAX model is same as the ARIMA model that we already got beforehand which is (p,d,q)=(2,1,1).

```
Regression with ARIMA(2,1,1) errors
Coefficients:
ar1      ar2      ma1    xreg
-1.2223  -0.7265  -1.0000  2e-04
s.e.    0.1057   0.1027   0.0742  2e-04
sigma^2 = 0.00456:  log likelihood = 46.63
AIC=-83.25   AICc=-81.38   BIC=-75.06
Training set error measures:
                      ME       RMSE        MAE       MPE      MAPE      MASE        ACF1
Training set -0.00444114 0.06304893 0.05064879 19.19142 73.35643 0.3720253 -0.4365445
```

The small AIC of the fitted model suggests the better fitting with the variable rainfall.

Now we have the forecast values that we have got from the training data and we already have our original test data values for yield with variables. So we now compare the two values and like previous method we include the trend and seasonlity in the data and repeat the process of model fitting for Yield with Annual Rainfall then we get the prediction based on the whole data which inlcudes the effect of exogenous variable : Annual Rainfall

The final Plot for Yield prediction with and without the effect of exogebous variable is given below:



The curve shows us that though the model fits well in theoritical approach as the AIC value is lower but there is nothing such serious effect of covariate is present in the data . This may be a fallacy of the observer or the researcher on the particular exogenous variable.

2.4.3 ARIMAX model for other different states

This kind of fallacious situation can arrive for forecasting the ARIMAX model for other State and for other crops also.

3. FINDINGS AND CONCLUSION

So far we conducted the Exploratory Data Analysis and time series forecasting with the Indian crop production data. We simply work with the yield data for different crops and different states with different time periods across the India.

The Exploratory analysis in the initial section helps us to compare the yield of different crops for different states with an interval of 10 years time period. It helps us to visualise the diversity in the season wise cultivation which includes summer , autumn, rabi, kharif etc. We use bubble plot for visualize the intensity of a crop in some different states.

Coming to the next section we have done a detailed study on time series fore casting for some of the major crops of the major agricultural states of India such as

From West Bengal we throroughly analyse the time series forecasting for the crops Rice and Jute; from Punjab we have studied the crops Wheat and Rice; from Uttar Pradesh we analysed the crops

Wheat and Bajra; from Madhya Pradesh we have analysed the crops Soyabeen and Groundnut; from Maharashtra we looked at the crops Sugarcane and Cotton.

We use ARIMA model for each crop prediction and checking the AIC of the models for getting a insight of model characteristics. In case of West Bengal we gain tremendous success as our model predicts the data for both rice and jute so well.

similar results shown for the states like Punjab ; there also in both the cases i.e wheat and rice prediction our model performs so well.

On the otherhand our model performs more or less good in the states like Maharashtra, Madhya Pradesh, Uttar Pradesh.

May be There are some other factors which effects the prediction.

In the third and final section of this thesis we were willing to test the effect of the several covariates of the model prediction like Annual Rainfall, Pesticides and Fertilizers . To analyse the effect of the agricultural covariates we introduced the concept of ARIMAX model . Here we were disappointed by the results of this model.
In case of prediction of Yield of rice in West Bengal, though there is significance in the effect of Annual Rainfall but in bare eyes it can't be visualize.

same fallacial situation arises for Yield production of Jute in West Bengal even in cae of other states also.

New models and better data sets are propoed to fit the better models and to gain the grater accuarcy.

4. REFERENCES:

- Time Series Forecasting ; Chris Chatfield; reader in Statistics, Department of Mathematical Sciences,University of Bath, UK
- **Introduction to Time Series and Forecasting;** Peter J. Brockwell , Richard A. Davis
- https://www.geeksforgeeks.org/bubble-plot-with-ggplot2-in-r/
- Time Series Forecasting with ARIMA , SARIMA and SARIMAX; https://towardsdatascience.com/time-series-forecasting-with-arima-sarima-and-sarimax-ee61099e78f6/
- FORECASTING FOR AGRICULTURAL PRODUCTION USING ARIMA MODEL; https://archives.palarch.nl/index.php/jae/article/download/5116/5043/9832
- https://upag.gov.in/