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

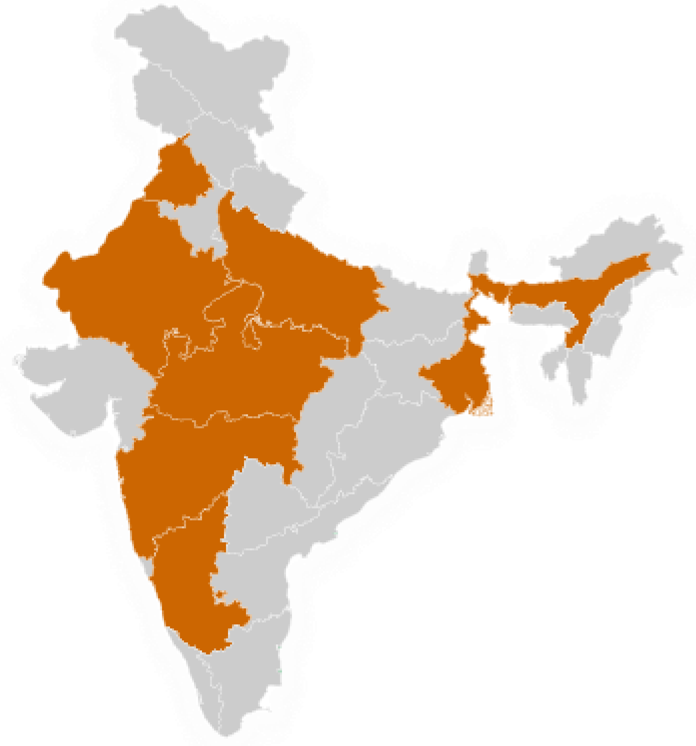
Prepared by: Ankita Pal & Nilanjan Barik
Class Roll Numbers: STAT 014, STAT 021
Registration Numbers: 23414270014, 23414110021

Agriculture in India

- The history of agriculture in India dates back to the **Neolithic Period**.
- India ranks **second in global farm outputs**. According to the Indian Economic Survey 2020 -21:
  - More than **55%** of the Indian workforce is employed in agriculture.
  - It contributes **20.2%** to the country's GDP.
  - Grew by **3.8%** in FY25, driven by record Kharif production, favourable monsoons and improved rural demand.
- India has the **world's largest** area planted for wheat, rice, and cotton, and is the largest producer of milk, pulses, and spices in the world.
- It is the **second-largest producer** of fruit, vegetables, tea, farmed fish, cotton, sugarcane, wheat, rice, cotton, and sugar.
- India has the **second-largest agricultural land** globally.

Key Agricultural Clusters in India

- Uttar Pradesh
- West Bengal
- Madhya Pradesh
- Karnataka
- Maharashtra
- Punjab
- Rajasthan
- Assam



Source: India Brand Equity Foundation (IBEF)
https://www.ibef.org/industry/agriculture-india

Objective

- Our aim is to use data analytics and statistical modeling to support agricultural forecasting and planning. With increasing climate variability and growing food security needs, predictive models are essential for guiding timely decisions on resource use, procurement, crop management and government policy-making.
- We were especially motivated by the chance to use time series analysis and ARIMA and ARIMAX models on real agricultural data. These tools helped us turn past crop yield records into useful insights. They're not only proven methods for forecasting but also easy to understand and flexible, which makes them ideal for agriculture.

Sample Snapshot of the Dataset

- Includes **crop yield** (production per unit area) data for **multiple crops** cultivated across **various states in India** from the year 1997 till 2019.
- Covers :
  - **crop types** (e.g. Rice, wheat, sugarcane)
  - **crop years** (1997-2019)
  - **cropping seasons** (e.g., Kharif, Rabi, Whole Year)
  - **area** under cultivation (in hectares)
  - **production** quantities (in metric tons)
  - **annual rainfall** (in mm)
  - **fertilizer** usage (in kilograms)
  - **pesticide** usage (in kilograms)

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

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

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
- xi: value at location i
- $\bar{x}$: mean of the variable
- wij: spatial weight

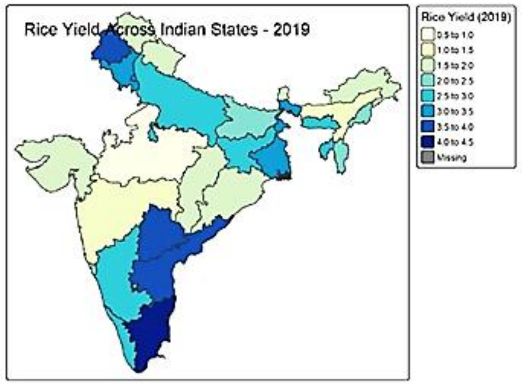
  between units i and j
- W: sum of all wij

Spatial weights to show how a state like Madhya Pradesh influences or is influenced by neighboring states like Uttar Pradesh, Maharashtra, etc. Example: Bihar and West Bengal are neighbors , weight = 1; Bihar and Kerala ,weight = 0.

Interpretation:

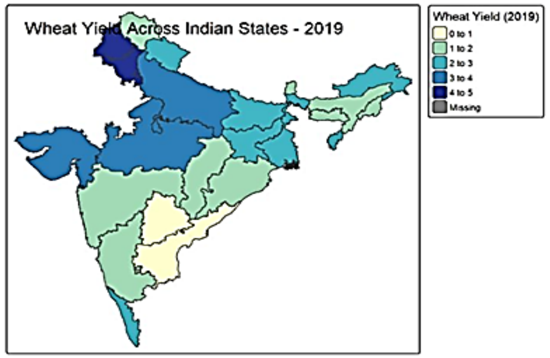
I>0: positive spatial autocorrelation (clusters)
I<0: negative spatial autocorrelation (dispersion)
I≈0: random pattern

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, parts of Maharashtra) had lower yields, in the 1.0 to 2.0 range. Central and eastern belt (e.g., Odisha) shows moderate productivity, around 2.5 to 3.0.

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.

As yield is measured as production per unit area, it is a unit-free measure. Therefore, for analysis purposes, we prefer using yield instead of production, especially since different crops are involved.

RICE

```
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
```

WHEAT

```
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
```

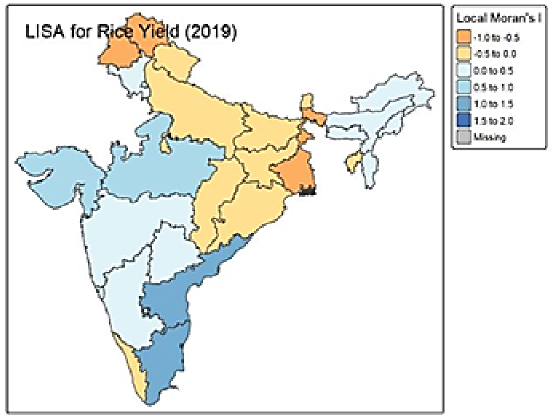
- **Moran's I for rice = 0.2025**: **Low positive spatial autocorrelation**
- **Moran's I for wheat = 0.5006** : **Moderate to high positive spatial autocorrelation**

- ❑ **Rice** is grown under highly localized conditions (e.g., irrigation, water availability, monsoon dependency). Its yield can vary significantly between states due to microclimatic and hydrological differences.
- ❑ Rice is typically grown in both **Kharif (monsoon)** and **Rabi (winter)** seasons in some states, while in others it's only in one season. This heterogeneity in cropping pattern affects spatial consistency.
- ❑ **Rice** production is more **dispersed** across a wide range of agro-climatic zones (e.g., West Bengal, Tamil Nadu, Assam), so weaker clustering is observed.

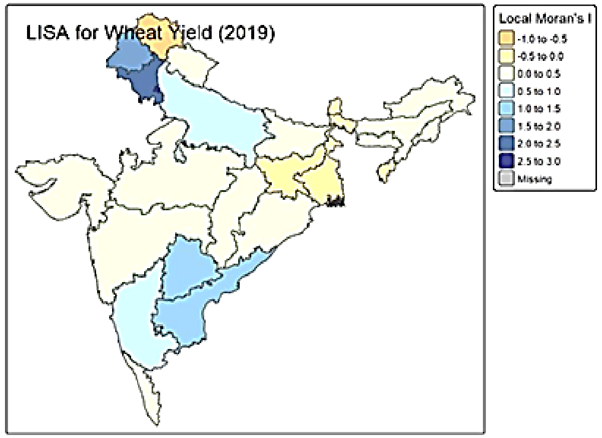
- ❑ **Wheat** is more uniformly cultivated across regions with similar agro-climatic zones ,leading to similar yields in neighboring regions.
- ❑ Wheat is mostly a **Rabi** crop with a consistent seasonality across states, contributing to higher regional coherence in yield.
- ❑ **Wheat** is more **concentrated** in specific adjacent regions (e.g., Punjab, Haryana, UP), leading to stronger spatial clustering.

❑ 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 analyze 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.



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.



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.

Checking for Stationarity

❑A time series is generally non-stationary. To make it stationary, we remove the trend and seasonality from the data. Then, we check for stationarity using the ADF(Augmented Dickey Fuller test) test .(By checking the p-value, if it is less than 0.05, we reject the null hypothesis at the 5% level of significance and conclude that the data is stationary).

❑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. The test statistic: $DF_\tau = \frac{\hat{\gamma}}{SE(\hat{\gamma})}$

❑Remove the trend and stationary part :

- When the dataset has a seasonal component, we use the **Decompose function** to remove both the trend and seasonality. If there is no seasonal component but a trend is present, we fit a suitable **polynomial trend equation** and remove the trend component from the data.
- Here if seasonality is constant over time we use additive model and if seasonality changes in proportion to the level of time series we use multiplicative model.

Time Series Models:

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 $\theta i \prime s (i = 1, 2, \cdots, q)$ 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:

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

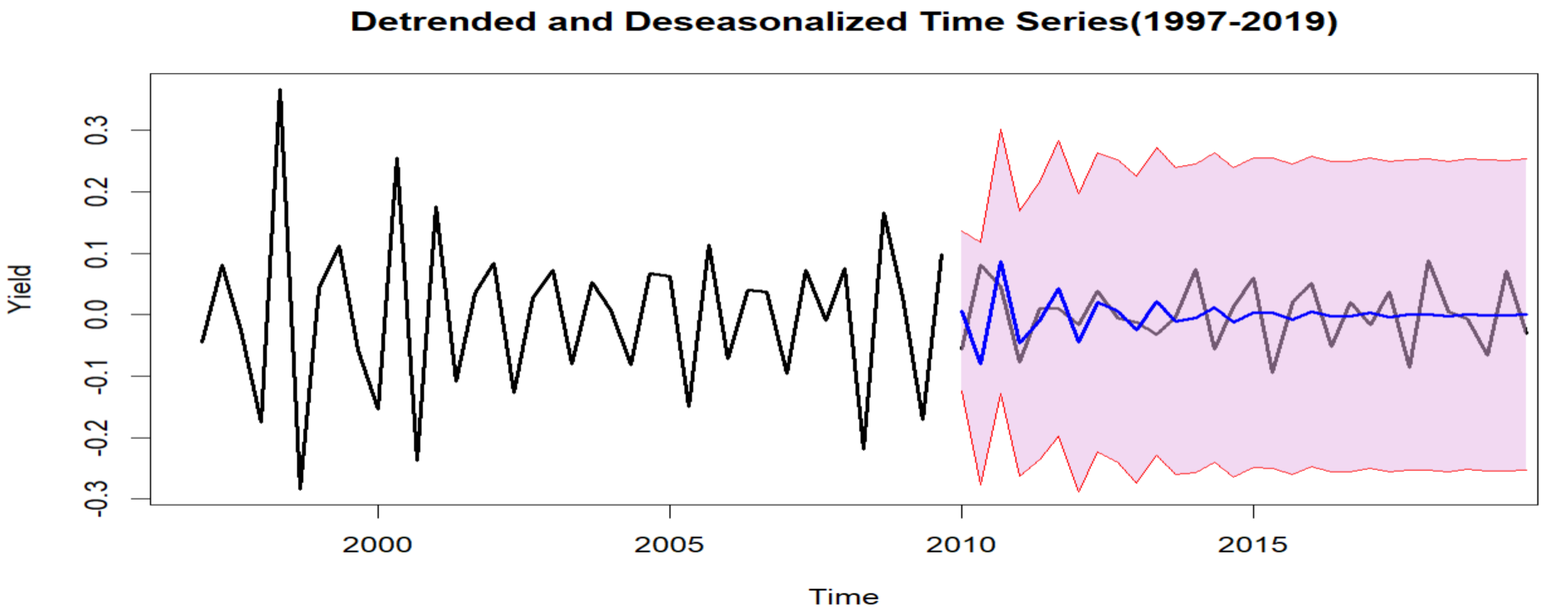
Where :

B is the backshift operator

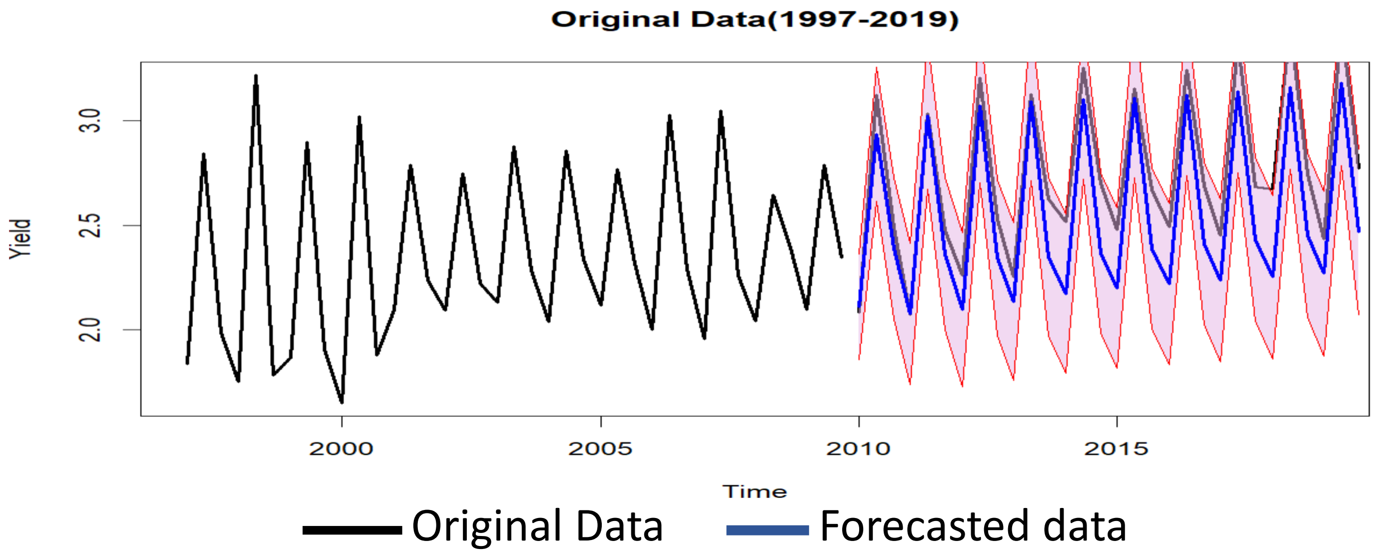
Forecasting Rice Yield in West Bengal (2020 to 2029):

Methodology:

- We first divide the dataset into two parts: *1997 to 2009* (training set) and *2010 to 2019* (test set). Both parts are converted into time series objects with a frequency of 3, corresponding to the three seasons: **Summer**, **Autumn**, and **Winter**. Using the training set, we forecast the values for the years 2010 to 2019 and calculate the Mean Squared Error (MSE) of the predictions using the test set.



	p	d	q	AIC	MSE
56	2	1	1	-83.548075	3.198841e-03
57	2	1	2	-103.029245	5.562386e-03
58	2	1	3	-103.159154	6.016969e-03
59	2	1	4	-105.158609	4.212202e-03
60	2	1	5	-103.843917	4.727510e-03
61	2	2	0	-10.534840	4.696196e-01
62	2	2	1	-42.155740	5.277707e-03

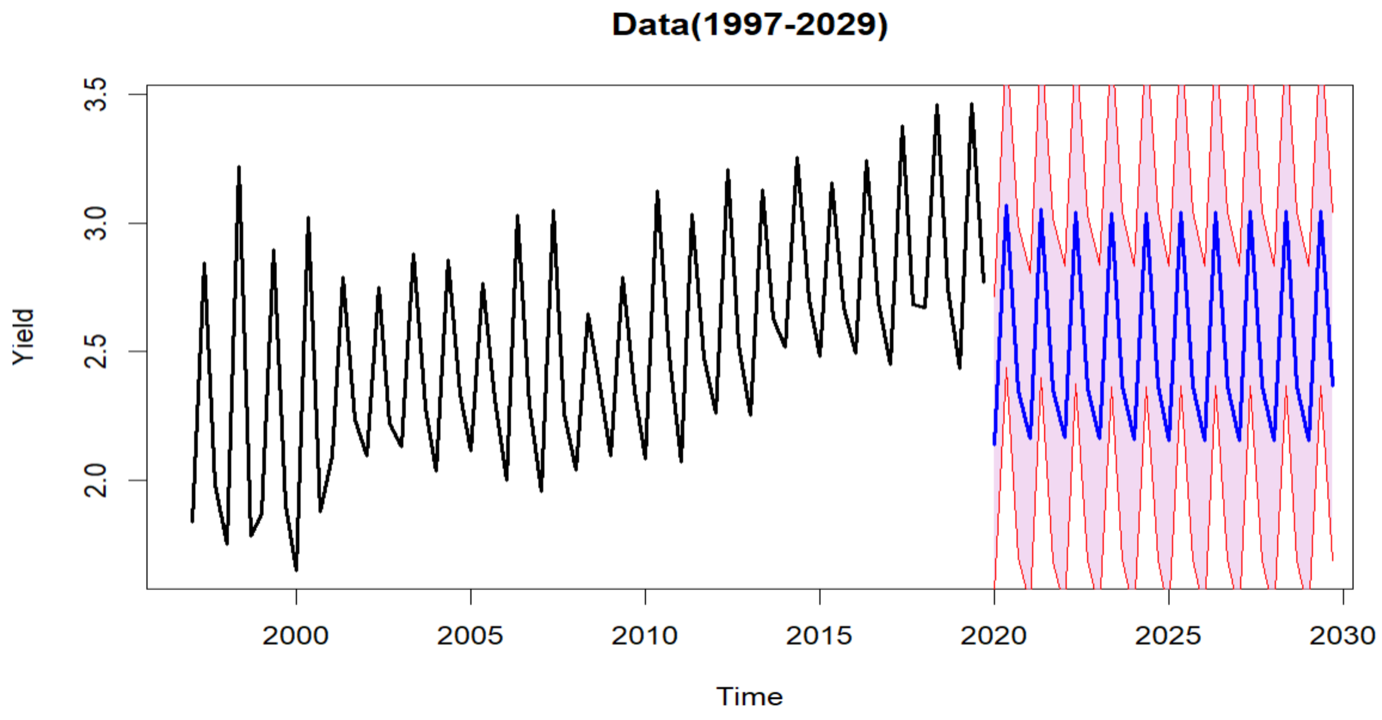


Black line indicates the original data (whole data i.e. yield from 1997 to 2019) blue line indicates the predicted data for the yield of jute from 2020 to 2029. Two red lines denotes the 95% confidence interval for the forecasted values .

- We then select the time series model that gives the **minimum MSE**. After selecting the best model, We forecast the future trend values using a simple linear trend equation($Y_t=a+bt$). Since there are only three seasonal values (0.3640174 0.5416963 -0.1776783) in the dataset, we repeat those values for the forecasted period as well. We **add back the trend and seasonal components** to the forecasted values and compare them with the actual data for the test years.

	Original_data	Forecasted_data	Error	Lower_95	Upper_95
1	2.081667	2.110555	-0.03	1.858276	2.362833
2	3.123889	2.936319	0.19	2.615698	3.256939
3	2.487222	2.389717	0.10	2.051076	2.728358
4	2.072222	2.075806	0.00	1.736406	2.415206
5	3.035000	3.024885	0.01	2.673773	3.375997
6	2.475556	2.364123	0.11	1.999320	2.728926
7	2.259444	2.095587	0.16	1.727248	2.463927
8	3.206111	3.073472	0.13	2.704415	3.442530
9	2.530000	2.346262	0.18	1.973258	2.719267
10	2.252778	2.135035	0.12	1.758485	2.511584
11	3.127222	3.093409	0.03	2.715498	3.471319
12	2.626111	2.347902	0.28	1.969201	2.726603
13	2.516111	2.172186	0.34	1.792004	2.552368
14	3.254444	3.102020	0.15	2.720269	3.483771
15	2.700556	2.364490	0.34	1.981947	2.747034
16	2.480000	2.199929	0.28	1.816639	2.583219
17	3.155000	3.110930	0.04	2.726583	3.495277
18	2.670556	2.387493	0.28	2.002241	2.772744
19	2.493810	2.219926	0.27	1.833863	2.605989
20	3.244286	3.124323	0.12	2.737439	3.511207
21	2.691364	2.410800	0.28	2.023068	2.798533
22	2.450000	2.236341	0.21	1.847717	2.624964
23	3.375714	3.141702	0.23	2.752249	3.531156
24	2.684545	2.432003	0.25	2.041704	2.822303
25	2.672381	2.252360	0.42	1.861162	2.643559
26	3.459524	3.161059	0.30	2.768983	3.553134
27	2.744545	2.451169	0.29	2.058199	2.844139
28	2.434286	2.269351	0.16	1.875469	2.663233
29	3.464286	3.180740	0.28	2.785942	3.575538
30	2.771364	2.469255	0.30	2.073517	2.864993

- Next, we take the full dataset, convert it into a time series object, and **detrend** and **deseasonalize** it. Using the previously selected ARIMA model (with fixed coefficients), we forecast the yield values for the **next 10 years**. Finally, we **add back the trend and seasonal components** to forecast the yield values corresponding to the original dataset.



	Point.Forecast	Lower_95	Upper_95
1	2.139839	1.566230	2.713448
2	3.070614	2.439827	3.701400
3	2.345283	1.699959	2.990608
4	2.160158	1.514645	2.805672
5	3.052783	2.397826	3.707739
6	2.351610	1.685819	3.017400
7	2.165551	1.497184	2.833918
8	3.041902	2.373405	3.710399
9	2.360507	1.689295	3.031719
10	2.162927	1.489197	2.836657
11	3.038586	2.364276	3.712895
12	2.366288	1.691877	3.040700
13	2.158524	1.483400	2.833648
14	3.039599	2.363773	3.715425
15	2.368265	1.692350	3.044180
16	2.155475	1.479531	2.831419
17	3.041758	2.365569	3.717946
18	2.367922	1.691590	3.044254
19	2.154325	1.477965	2.830685
20	3.043354	2.366971	3.719737
21	2.366874	1.690442	3.043306
22	2.154408	1.477924	2.830891
23	3.044011	2.367527	3.720495
24	2.366044	1.689558	3.042530
25	2.154910	1.478399	2.831421
26	3.044018	2.367502	3.720533
27	2.365674	1.689157	3.042192
28	2.155339	1.478817	2.831861
29	3.043780	2.367257	3.720303
30	2.365645	1.689118	3.042173

Forecasting Jute Yield in West Bengal (2020 to 2029):

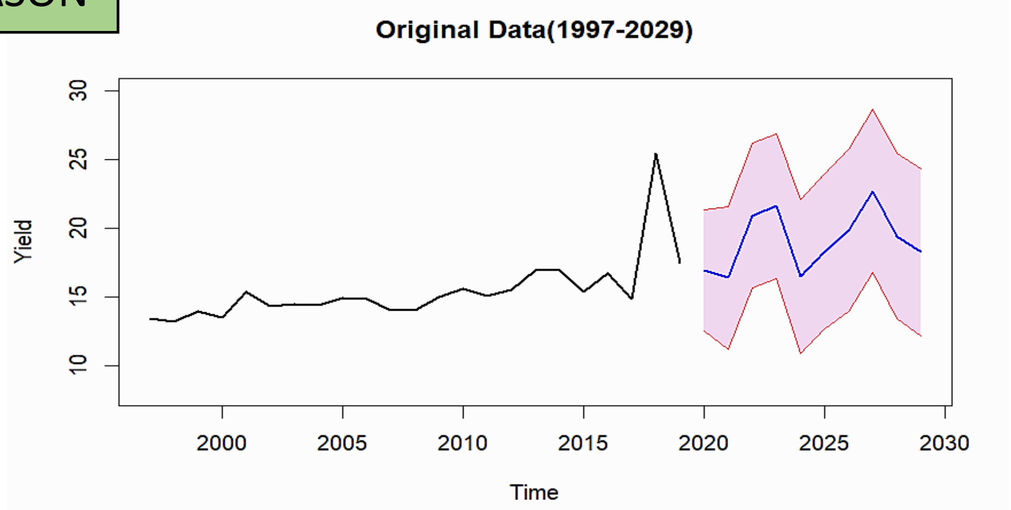
West Bengal is known for its jute production, particularly along the border with Bangladesh and south of the Ganges River. It is the 2nd largest crop in West Bengal in terms of yield. After rice it is the second most important exporting product that boost up the State's economy.

Here, we proceed in the same way as we did for rice yield forecasting. However, to make the dataset stationary, we first fit suitable polynomial trend equations and then remove the seasonality. (We create a loop to select the minimum degree of the polynomial trend equation such that, after removing the trend from the data, the resulting series becomes stationary.)

Jute yield in West Bengal was stable but spiked in 2019.That spike may be unnatural or misleading.

KHARIF SEASON

(MODEL:AR(5))

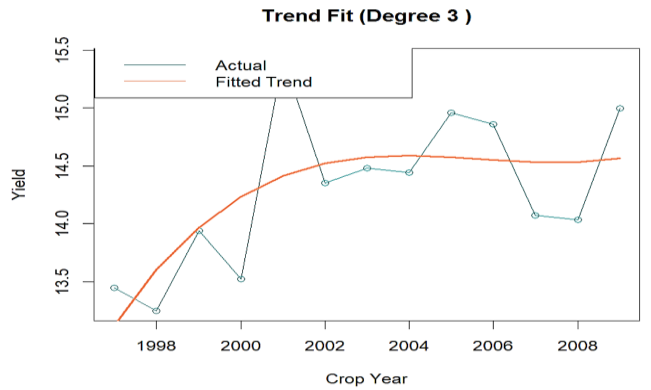


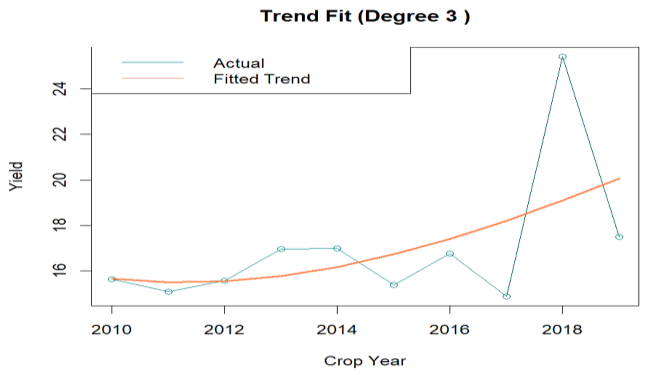
ORIGINAL DATA

Crop_Year	Yield
1997	13.44588
1998	13.24706
1999	13.94059
2000	13.52412
2001	15.42706
2002	14.35235

FORECASTED DATA

Point.Forecast	Lower_95	Upper_95
16.94393	12.55998	21.32789
16.39780	11.19319	21.60242
20.92637	15.70488	26.14786
21.60563	16.34470	26.86655
16.48583	10.89588	22.07579
18.32472	12.71595	23.93349
19.83904	13.94536	25.73272
22.69207	16.76805	28.61610
19.43635	13.46593	25.40677
18.25586	12.17894	24.33278

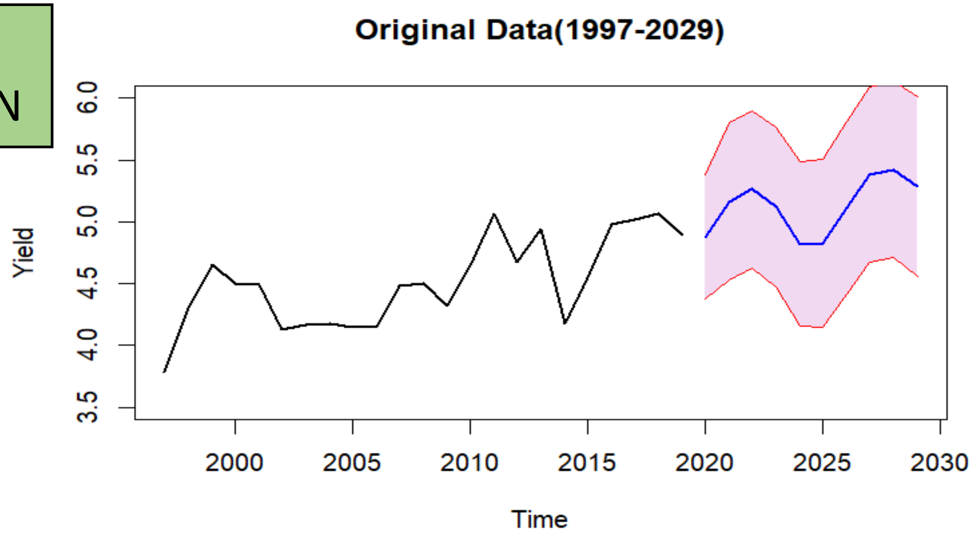




Punjab

WHEAT (MODEL : ARMA(6,3))

RABI SEASON



ORIGINAL DATA

Crop_Year	Yield
1997	3.781176
1998	4.298824
1999	4.651765
2000	4.503529
2001	4.487059
2002	4.130000

FORECASTED DATA

Point.Forecast	Lower_95	Upper_95
4.876710	4.372452	5.380969
5.165354	4.530317	5.800392
5.262053	4.621964	5.902142
5.119499	4.470783	5.768216
4.818053	4.153833	5.482274
4.825748	4.143571	5.507924
5.112155	4.415465	5.808845
5.379495	4.669373	6.089617
5.423604	4.708318	6.138889
5.284993	4.559722	6.010265

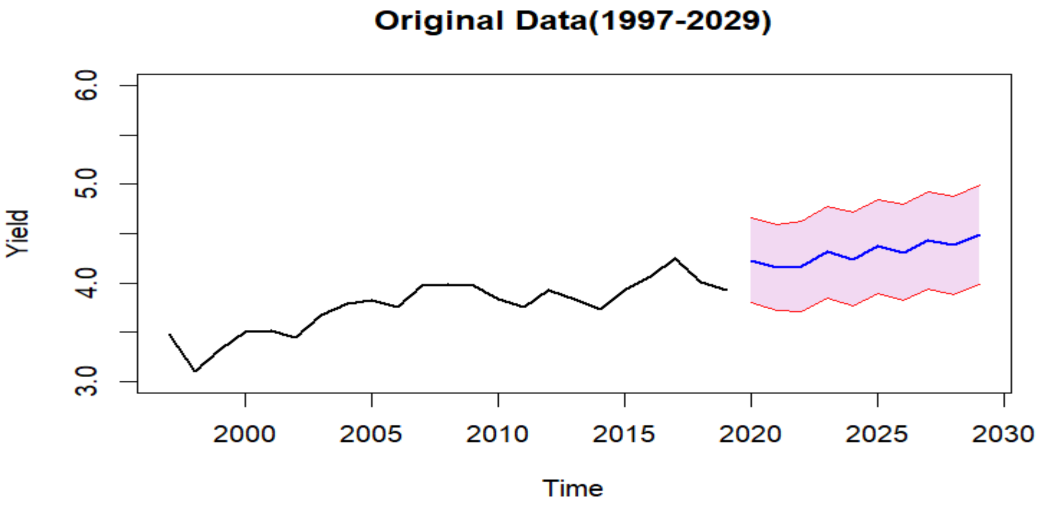
The wheat yield has been rising over the years, and the forecast shows increase in yield.

Possible reasons:

- ❑ Strong government support for wheat (like Minimum Support Price).
- ❑ Good irrigation systems in Punjab, especially for the Rabi season.
- ❑ Use of modern farming tools and machines that help improve wheat productivity. Wheat being less affected by monsoon, so its yield is more stable.

KHARIF SEASON

RICE (MODEL:ARIMA(1,1,4))



Rice yield has been slowly increasing, but the rise is more gradual and stable.

Possible reason:

- ❑ Rice depends more on the monsoon, so bad rainfall years affect yield.

ORIGINAL DATA

Crop_Year	Yield
1997	3.476471
1998	3.099412
1999	3.332353
2000	3.501765
2001	3.517059
2002	3.438824

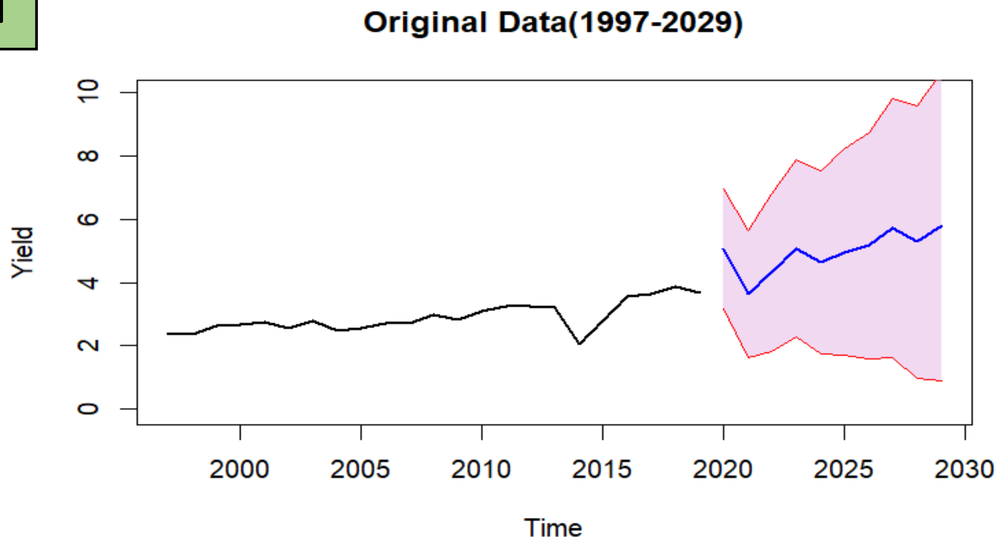
FORECASTED DATA

Point.Forecast	Lower_95	Upper_95
4.227987	3.795462	4.660513
4.158087	3.722749	4.593425
4.166933	3.706453	4.627414
4.313679	3.850144	4.777214
4.239128	3.764427	4.713829
4.371562	3.893506	4.849618
4.310396	3.822096	4.798697
4.430310	3.938306	4.922314
4.380855	3.879261	4.882450
4.489816	3.984152	4.995480

Uttar Pradesh

RABI SEASON

WHEAT (MODEL:ARIMA(5,2,0))



ORIGINAL DATA

Crop_Year	Yield
1997	2.416000
1998	2.373735
1999	2.636420
2000	2.680580
2001	2.729565
2002	2.558551

FORECASTED DATA

Point.Forecast	Lower_95	Upper_95
5.066140	3.1581425	6.974137
3.637234	1.6210936	5.653374
4.316570	1.8341059	6.799034
5.075571	2.2822054	7.868937
4.631556	1.7285658	7.534546
4.957578	1.6970332	8.218122
5.163874	1.5720719	8.755675
5.727393	1.6378749	9.816911
5.284779	0.9760617	9.593495
5.788654	0.9084101	10.668898

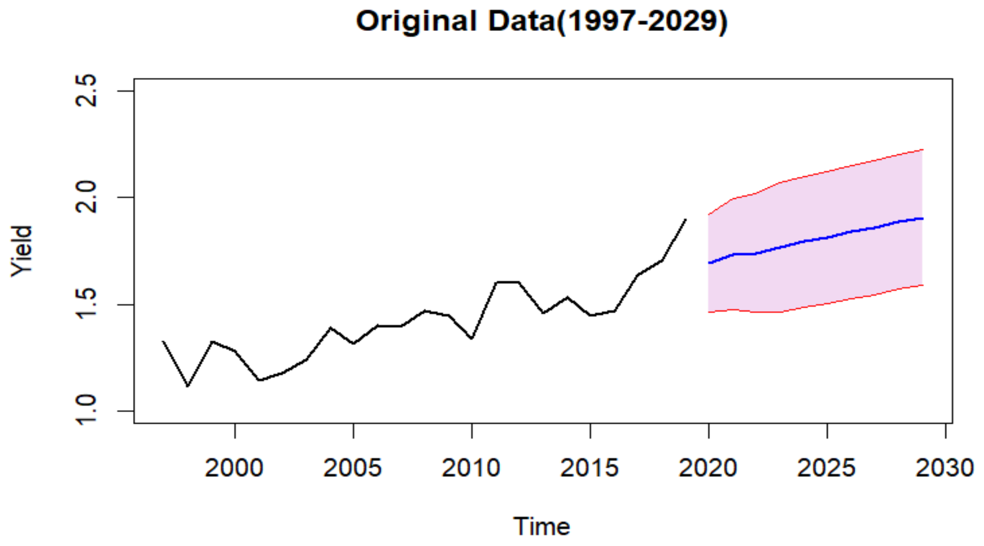
Wheat yield increased slightly over time.

Possible reason:

- The prediction shows a wider uncertainty range, meaning future wheat yields may go up or down depending on extreme weather events (heatwaves, unseasonal rains).

KHARIF SEASON

BAJRA (MODEL:ARMA(1,5))



ORIGINAL DATA

Crop_Year	Yield
1997	1.324000
1998	1.110714
1999	1.326176
2000	1.278824
2001	1.143333
2002	1.176324

FORECASTED DATA

Point.Forecast	Lower_95	Upper_95
1.693912	1.465179	1.922645
1.734613	1.474783	1.994443
1.739517	1.460667	2.018366
1.767768	1.462212	2.073323
1.794652	1.486685	2.102619
1.813524	1.503397	2.123651
1.840435	1.528058	2.152812
1.860514	1.545797	2.175230
1.886400	1.569257	2.203544
1.907350	1.587694	2.227006

Bajra yield is expected to grow strongly.

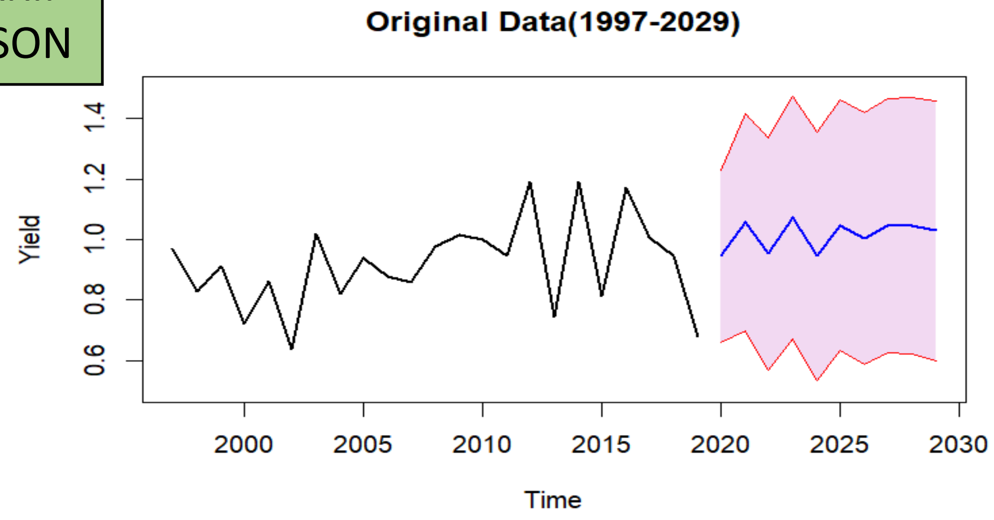
Possible reasons:

- ❑ Bajra is a climate-resilient crop that does well even in dry and hot conditions, which is useful for Kharif season.
- ❑ The government is now promoting millets like bajra under the "International Year of Millets 2023" and other nutrition-based schemes.

Madhya Pradesh

KHARIF SEASON

SOYABEAN (MODEL:AR(6))



ORIGINAL DATA

Crop_Year	Yield
1997	0.9697727
1998	0.8259615
1999	0.9103509
2000	0.7197778
2001	0.8622222
2002	0.6355556

FORECASTED DATA

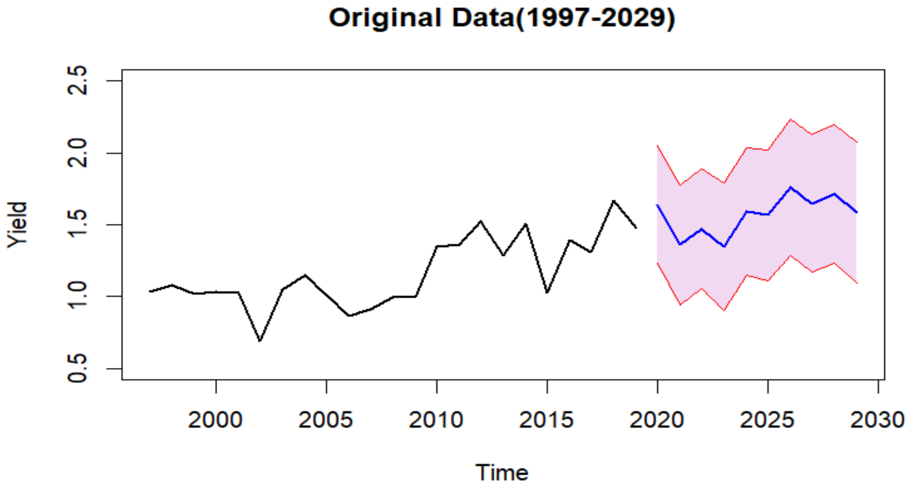
Point.Forecast	Lower_95	Upper_95
0.9450066	0.6605954	1.229418
1.0570579	0.6978396	1.416276
0.9533641	0.5684779	1.338250
1.0726263	0.6722520	1.473001
0.9455306	0.5341471	1.356914
1.0474568	0.6328781	1.462036
1.0029501	0.5854150	1.420485
1.0468701	0.6261149	1.467625
1.0460683	0.6213102	1.470826
1.0287554	0.5986709	1.458840

Soyabean shows high yield variability. Madhya Pradesh is India's largest soyabean producer, so any fluctuations matter nationally.

The forecast shows moderate increase in yield, but the wide prediction band indicates future yield is uncertain.

GROUNDNUT (MODEL:AR(5))

KHARIF SEASON



Groundnut yield is steadier than soyabean.

Possible reasons:

- ❑ Better seed varieties.
- ❑ Less water-intensive than other oilseeds

ORIGINAL DATA

Crop_Year	Yield
1997	1.0311429
1998	1.0804348
1999	1.0146667
2000	1.0302222
2001	1.0233333
2002	0.6888889

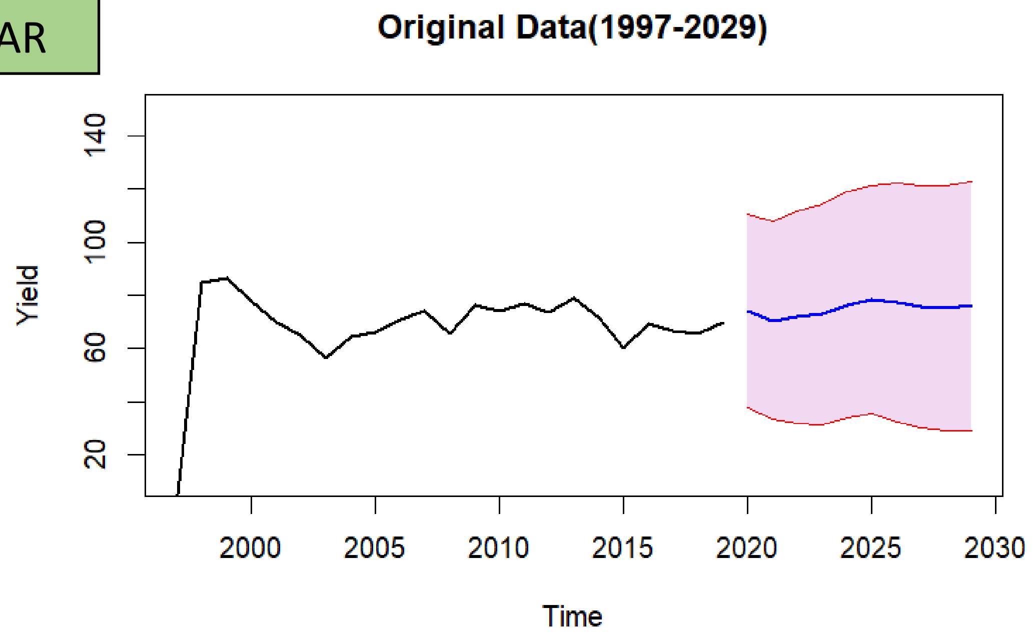
FORECASTED DATA

Point.Forecast	Lower_95	Upper_95
1.590530	1.1869354	1.994125
1.355683	0.9400526	1.771312
1.483177	1.0627703	1.903583
1.431285	1.0002830	1.862288
1.581679	1.1480115	2.015346
1.513705	1.0703511	1.957058
1.675211	1.2127572	2.137664
1.588408	1.1220489	2.054768
1.655803	1.1848854	2.126721
1.592212	1.1152069	2.069217

Maharashtra

WHOLE YEAR

SUGARCANE (MODEL:AR(4))



Sugarcane yields went up sharply around 1998, then dropped, then moved up and down but kind of settled over the years.

Possible reasons:

- ❑ Sugarcane needs rich, well-irrigated soil , not available everywhere.
- ❑ Sugarcane needs a long, warm growing season

ORIGINAL DATA

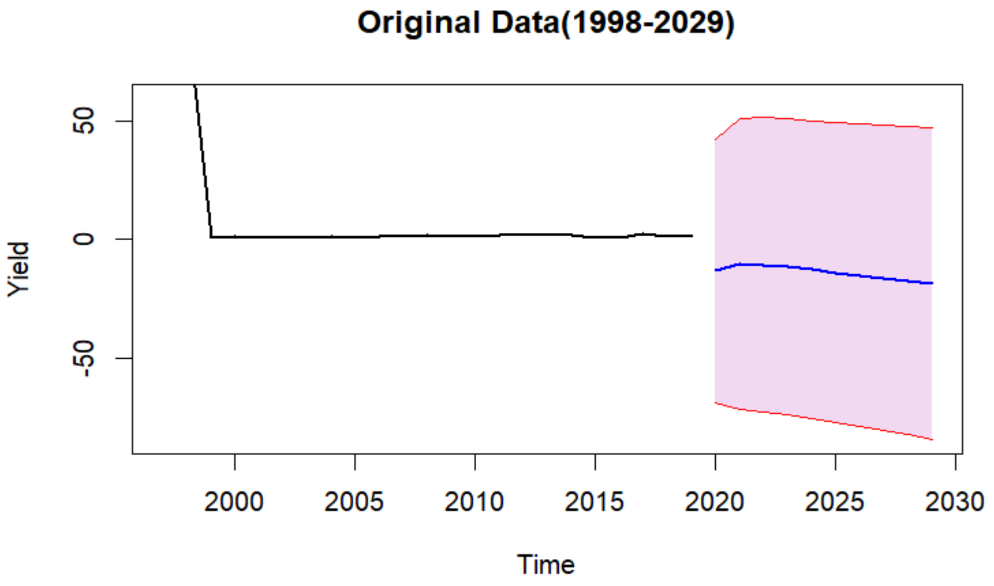
Crop_Year	Yield
1997	0.8633333
1998	84.5543478
1999	86.5557692
2000	77.6540741
2001	69.5774074
2002	65.2466667

FORECASTED DATA

Point.Forecast	Lower_95	Upper_95
74.10966	37.50599	110.7133
70.61766	33.49620	107.7391
71.77941	31.76041	111.7984
72.87194	31.30252	114.4414
76.37197	33.89652	118.8474
78.64062	35.76987	121.5114
77.43534	32.45249	122.4182
75.84219	30.28912	121.3953
75.18442	29.34404	121.0248
76.06072	29.34026	122.7812

KHARIF SEASON

COTTON (MODEL:ARMA(1,1))



Cotton is an economically significant crop in Maharashtra. Maharashtra is India's largest producer of cotton (by area).

There is a flat zero yield line from 2000 to 2018. The forecast (2020–2029) starts below zero, which is not possible for yield. The confidence interval is very wide and symmetric, indicating high uncertainty. So it shows clear signs of data issues.

ORIGINAL DATA

Crop_Year	Yield
1997	95.9939130
1998	1.0404348
1999	1.2138462
2000	0.8811538
2001	0.8973077
2002	0.9903704

FORECASTED DATA

Point.Forecast	Lower_95	Upper_95
-13.26906	-69.07727	42.53915
-10.32251	-71.94791	51.30289
-10.62040	-72.87435	51.63355
-11.60694	-74.30571	51.09183
-12.73965	-75.89899	50.41969
-13.90339	-77.54707	49.74029
-15.07371	-79.22495	49.07753
-16.24543	-80.92651	48.43564
-17.41745	-82.64970	47.81480
-18.58953	-84.39337	47.21431

ARIMAX model(Auto Regressive Integrated Moving Average with Exogenous Variables)

Effect of covariates on prediction

A Modification to the ARIMA Model:

In the ARIMA model, we predict yield based on the yield data already available. However, if we want to test whether our predictions are influenced by any exogenous variables, such as Annual Rainfall, Fertilizer usage, or Pesticide application, a better approach is to upgrade the ARIMA model to an **ARIMAX** model.

ARIMAX Model:

An ARIMAX model, which stands for *Autoregressive Integrated Moving Average with Exogenous inputs*, is an enhanced version of the ARIMA (Autoregressive Integrated Moving Average) model. The ARIMAX model extends the ARIMA model by incorporating exogenous variables, external factors that may influence the time series under study. This integration enables the model to utilize additional information, improving the forecasting accuracy significantly.

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

where $\boldsymbol{Y_t}$ is the value of the dependent variable at time **t**, $\boldsymbol{\phi_i}$ are the parameters of the autoregressive part, $\boldsymbol{\theta_j}$ are the parameters of the moving average part, $\boldsymbol{x_t}$ represents the exogenous variables at time **t**, and $\boldsymbol{\beta}$ are the coefficients associated with these exogenous variables.

Rice Prediction in West Bengal

As mentioned earlier, the **training dataset includes the yield of rice in West Bengal from 1997 to 2009**, along with each of the exogenous variables. The **test set includes the yield data from 2010 to 2019**, along with the corresponding exogenous variables.

For each exogenous variable:

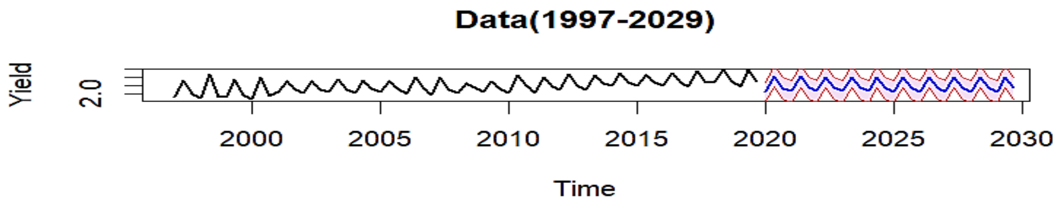
- We perform **decomposition** to separate and remove the **trend and seasonal** components from the training data. Then perform **the Augmented Dickey-Fuller (ADF) test** to check stationarity.
- This process results in a **stationary version of each exogenous variable**.
- We then proceed to fit the **ARIMAX model using Annual Rainfall** as the exogenous variable for the training dataset.
- The **order of the ARIMAX model is kept the same as that of the ARIMA model** previously identified, which is **(p, d, q) = (2, 1, 1)**.

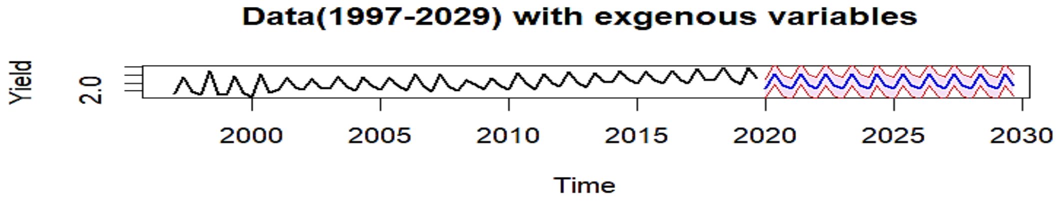
Now, we have the forecasted yield values obtained from the training data, and we also have the actual yield values from the test data. So, we compare these two. As in the previous method, we add the trend and seasonality back into the data and repeat the yield forecasting process .Then we predict the yield for next 10 years , including Annual Rainfall as an exogenous variable. This gives us predictions based on the full data, capturing the effect of Annual Rainfall.

```
Regression with ARIMA(2,1,1) errors

Coefficients:
          ar1      ar2      ma1  new$Annual_Rainfall  new$Pesticide  new$Fertilizer
      -1.3168  -0.7106  -1.0000                2e-04              0               0
s.e.   0.0456   0.0598   0.0878                2e-04            NaN             NaN

sigma^2 = 0.004136:  log likelihood = 49.55
AIC=-85.1   AICc=-81.37   BIC=-73.64

Training set error measures:
                       ME       RMSE        MAE      MPE     MAPE      MASE      ACF1
Training set -0.009344398 0.05825495 0.04779739 13.20715 79.36161 0.3510812 -0.356439
```

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

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





	Point.Forecast	Lower_95	Upper_95
1	2.139839	1.566230	2.713448
2	3.070614	2.439827	3.701400
3	2.345283	1.699959	2.990608
4	2.160158	1.514645	2.805672
5	3.052783	2.397826	3.707739
6	2.351610	1.685819	3.017400
7	2.165551	1.497184	2.833918
8	3.041902	2.373405	3.710399
9	2.360507	1.689295	3.031719
10	2.162927	1.489197	2.836657
11	3.038586	2.364276	3.712895
12	2.366288	1.691877	3.040700
13	2.158524	1.483400	2.833648
14	3.039599	2.363773	3.715425
15	2.368265	1.692350	3.044180
16	2.155475	1.479531	2.831419
17	3.041758	2.365569	3.717946
18	2.367922	1.691590	3.044254
19	2.154325	1.477965	2.830685
20	3.043354	2.366971	3.719737
21	2.366874	1.690442	3.043306
22	2.154408	1.477924	2.830891
23	3.044011	2.367527	3.720495
24	2.366044	1.689558	3.042530
25	2.154910	1.478399	2.831421
26	3.044018	2.367502	3.720533
27	2.365674	1.689157	3.042192
28	2.155339	1.478817	2.831861
29	3.043780	2.367257	3.720303
30	2.365645	1.689118	3.042173

	Point.Forecast_exo	Lower_95	Upper_95
1	2.140538	1.563759	2.713448
2	3.069677	2.436203	3.701400
3	2.347080	1.700342	2.990608
4	2.158330	1.511203	2.805672
5	3.053044	2.396305	3.707739
6	2.352719	1.686179	3.017400
7	2.163789	1.495381	2.833918
8	3.043061	2.374388	3.710399
9	2.360478	1.689161	3.031719
10	2.161964	1.488640	2.836657
11	3.039784	2.366137	3.712895
12	2.365567	1.691745	3.040700
13	2.158442	1.483980	2.833648
14	3.040278	2.365321	3.715425
15	2.367489	1.692505	3.044180
16	2.155892	1.480859	2.831419
17	3.041864	2.366626	3.717946
18	2.367465	1.692148	3.044254
19	2.154805	1.479481	2.830685
20	3.043132	2.367776	3.719737
21	2.366775	1.691385	3.043306
22	2.154706	1.479288	2.830891
23	3.043732	2.368315	3.720495
24	2.366157	1.690737	3.042530
25	2.154995	1.479557	2.831421
26	3.043836	2.368398	3.720533
27	2.365835	1.690398	3.042192
28	2.155291	1.479848	2.831861
29	3.043722	2.368279	3.720303
30	2.365758	1.690313	3.042173

The curve shows that although the model fits well from a theoretical perspective, as indicated by the lower AIC value, there is no significant effect of the covariate present in the data. This may be due to a misinterpretation by the observer or researcher regarding the influence of the particular exogenous variable.

Findings and Conclusions:

- **Scope of Analysis**:

•A detailed study was conducted on time series forecasting for major crops across major agricultural states of India.

- The states and corresponding crops studied are:
    - **West Bengal**: Rice and Jute
    - **Punjab**: Wheat and Rice
    - **Uttar Pradesh**: Wheat and Bajra
    - **Madhya Pradesh**: Soyabean and Groundnut
    - **Maharashtra**: Sugarcane and Cotton

- **Model Used**:
    - For each crop, the **ARIMA model** was employed for yield prediction.

- **ARIMAX Model**:
  - •In the final section, we tested the influence of agricultural covariates: **Annual Rainfall, Pesticides, and Fertilizers**.
  - •To analyze their effect, we extended the ARIMA model to an **ARIMAX model**.
- **Results from ARIMAX**:
  - •Unfortunately, the ARIMAX model did not perform as expected.
  - •**West Bengal (Rice)**:
    Though Annual Rainfall showed statistical significance, its visual effect on prediction was negligible.
  - •**West Bengal (Jute)**:
    Similar lack of visible influence from covariates.
- **Conclusion and Future Work**:
  - •The disappointing results from the ARIMAX model highlight the need for:
  - •Improved datasets with more detailed and accurate covariate information.
  - •Exploration of **alternative or more complex models** that can better capture the dynamics between crop yield and influencing factors.
- Future studies may focus on integrating **spatial, climatic, and socio-economic variables** to enhance model performance and interpretability.