ASUTOSH COLLEGE

Estd. 1916

92, Shyama Prasad Mukherjee Rd, Kolkata - 700026 (Affiliated to University of Calcutta)

DEPARTMENT OF STATISTICS

NAME : SOHINI BHADRA

STREAM : B.Sc. HONOURS

COLLEGE ROLL NO. : 91

UNIVERSITY ROLL NO. : 193012-11-0287

UNIVERSITY REGISTRATION NO. : 012-1211-0539-19

SUBJECT : STATISTICS
PAPER : DSE-B2 (PROJECT WORK)

SEMESTER : VI

YEAR : 2022

FOOD-CROPS PRODUCTION AND ECONOMIC GROWTH IN INDIA – A CASE STUDY AND DATA ANALYSIS

Contents

Executive summary of the project

Agriculture continues to be the backbone of the Indian economy - this is hardly an extravagant statement. India is the world's second most populous country after China. India is the world's largest producer of jute, pulses, and milk and ranks as the second largest producer of wheat, rice, cotton, groundnut, sugarcane, and horticulture crops. India ranks second worldwide in farm outputs. As per 2018, agriculture employed more than 50% of the Indian work force and contributed 17–18% to country's GDP.

So here we have collected data for four major food-grains produced in India and also the data on GDP over the last 50 years 1971-2020. Firstly, we analyse the individual production data of the four major food-grains using divided bar diagram, column charts. Secondly, we analyse the total production based on these four crops and GDP using scatter plot.

Here the collected data is clearly time series data (A time series may be defined as a collection of random variables which are ordered in time and defined on a set of time points.).We know the main objective in analysing time series is to understand , interpret and evaluate changes in economic phenomena in the hope of more correctly anticipating the course of future events.

So, now we have used time series analysis method to study the underlying pattern of rice, wheat , pulses and coarse cereals production of whole India over time and analyse the trend. Then we have used time series modelling to fit a suitable trend line equation for total production data (based on four major food-grains) and tried to predict the future production of upcoming years through the estimated trend line equation.

In the pre COVID-19 pandemic has disrupted the Indian agricultural system extensively. Nevertheless, the recent quarterly GDP estimates post-COVID scenario showcase robustness and resilience in Indian agriculture, the only sector to register a positive growth of 3.4% during the financial year (FY here after) 2020–21 (Quarter 1: April 2020 to June 2020). At the same time, the immediate past quarter growth was estimated at 5.9% witnessing a decline by 2.5% point. In this context, we aim to synthesize the early evidence of the COVID-19 impact on the Indian agricultural system viz., production, marketing and consumption followed by a set of potential strategies to recover and prosper post-pandemic. Survey findings indicate that the pandemic has affected production and marketing through labour and logistical constraints, while the negative income shock restricted access to markets and increased prices of food commodities affecting the consumption pattern. The pandemic wreaked a substantial physical, social, economic and emotional havoc on all the stakeholders of Indian agricultural system.

In our present study, the inter-relationships between agricultural production and economic growth are studied in an interdependent framework for economic growth keeping in view set policy guidelines for agricultural development in India using multiple linear regression growth model. The econometric model is formulated with aggregate information available over the period of 1971 - 2020 (economic reforms to new India concept) and both static and Augmented Dickey - Fuller test (ADF) along with Johansen co-integration test and regression analysis are carried out to assess the performance of the set model. This study examines the agricultural production and its impact on economic growth in India. The study reveals that if there had been no increase in agricultural output (explanatory variables) in India, it would have negatively impacted the economic growth in India.

Introduction

AGRICULTURE CONTINUES TO BE THE BACKBONE OF THE INDIAN ECONOMY - THIS IS HARDLY AN EXTRAVAGANT STATEMENT.

With a population of 1.27 billion, India is the world's second most populous country after China. Over two thirds of the country's population is directly dependent on agriculture. India is the world's largest producer of milk, pulses, and jute and ranks as the second largest producer of wheat, rice, cotton, groundnut, sugarcane, and horticulture crops. Agriculture and its allied sectors are the largest sources of livelihood, and these account for 23% of the GDP and employed 59% of the country's total workforce in 2016. According to World Bank Data 2018, India is the sixth largest economy by nominal GDP and third largest by purchasing power parity (PPP).

In more developing countries (MDC) like India, the agricultural sector, its growth, and agricultural production have been regarded as dominant prerequisites for economic growth. Several studies have been conducted, and we have considered the agricultural production as a dominant prerequisite. Since 1947, the agricultural sector has witnessed huge technological reforms, however, despite this fact, the farmers' standards of living did not improve (Bandaru, 2019). The government has also implemented the policy of regulated agricultural market, and a study on such a regulated agricultural market by Rehman (2015) found that the farmers had a positive perception towards the regulated markets. The idea behind implementing all the reforms was to increase agricultural production so as to make the country self-sufficient in agricultural production and also to generate employment opportunities in rural areas and increase the rural development so that the whole process will lead to economic growth. As per available data on rural development of India, it is essential to increase the agricultural production and generate employment opportunities in rural India. It will mean more adequate food supplies and will reduce the imports of the food items in the country. Moreover, when agricultural production will be in surplus, more foreign currency will be available for the development of the industrial and services sectors of the country. Inevitably, India's economic growth will increase. Consequently, the major improvements in the performance of the Indian economy can be initiated by agricultural development.

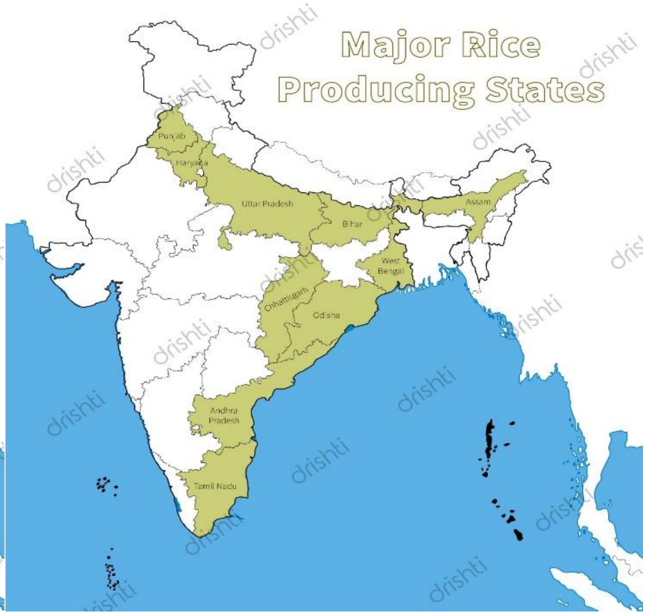
In the present study, we have attempted to ascertain the relationship of agriculture production with economic growth since 1971 - 2020 in India. The study is primarily based on the assumption that agriculture is of utmost importance in the Indian economy.
Major crops can be classified into-

Food crops- Rice, Wheat, Pulses and coarse cereals (sorghum, pearl millet,ragi, small millets,maize and barley)

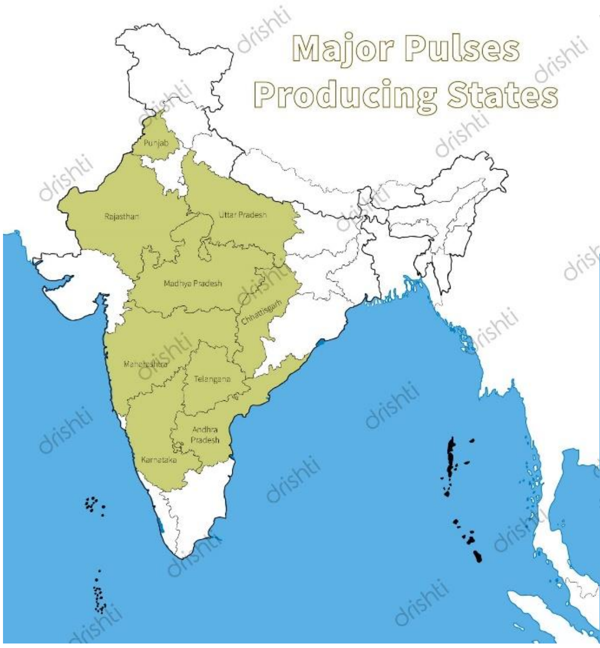
Cash crops- Sugarcane, Oilseeds, Horticulture crops, Tea, Coffee, Rubber, Cotton and Jute . Here we work with major food crops.

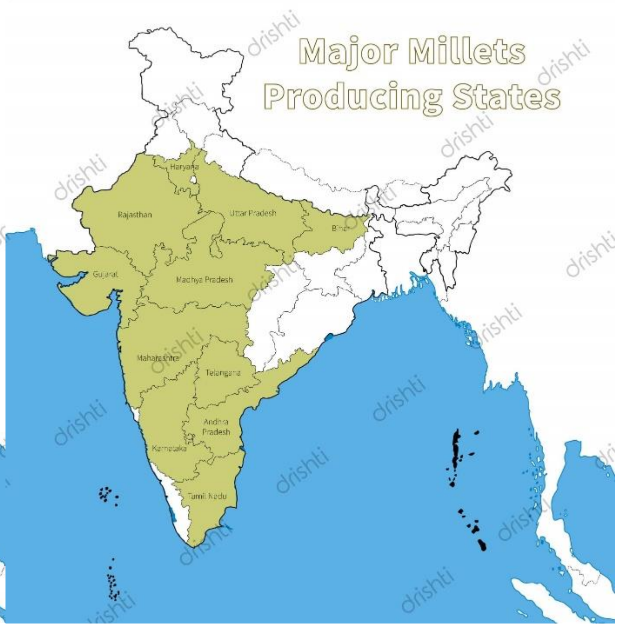
Wheat: Wheat is a major food crop in India. In 2019 - 20, Uttar Pradesh was the largest wheat producing states in India. Uttar Pradesh produced 33815.5 million tonnes of wheat in 2019-20 and Madhya Pradesh was the second largest wheat producing states in India. Madhya Pradesh produced 19607.1 million tonnes of wheat in the same year.

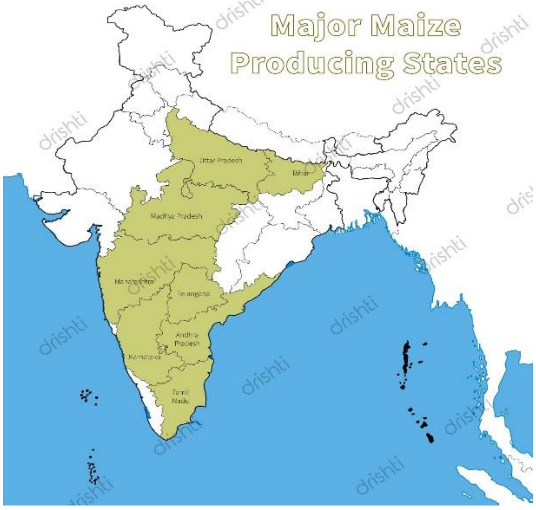
Rice : India is the largest rice producing country in the world after China. Rice contributes to more than 40% of the country's total food grain production. In India, rice is grown in 43.86 million ha, the production level is 104.8million tonnes, and the productivity is about 2390 kg/ha (Ministry of Agriculture & Farmers Welfare, Department of Agriculture, Cooperation & Farmers Welfare, Directorate of Economics and Statistics, 2016).



Pulses : Pulses are an important commodity, are rich in proteins, and are leguminous crops. Pulses being rich in protein complement the cereals substantially for the pre-dominantly vegetarian population of the country. Pulses can be produced with a minimum use of resources and are mostly cultivated in rainfed areas as these do not require intensive irrigation facility. In comparison with other crops and vegetables, pulses are grown in the fields left after satisfying the demand for cereal/cash crops.







coarse cereals: A variety of coarse cereals are grown throughout the country in different ecology, agroclimatic condition, but mostly as rainfed crop. Sorghum, pearl millet, maize, barley, finger millet and several small millets such as kodo millet, little millet, foxtail millet, proso millet and barnyard millet together called coarse cereals. Sorghum, pearl millet, finger millet, maize and small millets (barnyard millet, proso millet, kodo millet and foxtail millet) are also called nutri-cereals. Globally, average production of coarse grains is estimated to be about 1130.25 million tonnes during 2007-2011 and India contributed 3.6% (40.19 million tonnes) in global production of coarse grains and India ranks 4th after USA, China & Brazil.

COLLECTED DATA

Collected data on annual production of rice,wheat, pulses and cereal coarses of India from 1971-2020 and as well the GDP of India for the same time period is given below in table 1.

Here we have collected the data of production at the year time point and the production's unit is lakh tonnes. The data on GDP is collected in mktp (current US $).

We will use this unit (lakh tonnes) of production further for studying or analysing in our project.

Table 1: agricultural production of food-grains						
						(Lakh tonnes)
Year	Cereals				Total foodgrains	Gdp mktp(current US$)
	Rice	Wheat	Coarse Cereals	Pulses		
1	2	3	4	5	6	7
1971	430.7	264.1	246.0	110.9	1051.7	67350988021
1972	392.4	247.4	231.4	99.1	970.3	71463193830
1973	440.5	217.8	288.3	100.1	1046.7	85515269586
1974	395.8	241.0	261.3	100.2	998.3	99525899116
1975	487.4	288.4	304.1	130.4	1210.3	98472796457
1976	419.2	290.1	288.8	113.6	1111.7	1.02717E+11
1977	526.7	317.5	300.2	119.7	1264.1	1.21487E+11
1978	537.7	355.1	304.4	121.8	1319.0	1.373E+11
1979	423.3	318.3	269.7	85.7	1097.0	1.52992E+11
1980	536.3	363.1	290.2	106.3	1295.9	1.86325E+11
1981	532.5	374.5	310.9	115.1	1333.0	1.93491E+11
1982	471.2	427.9	277.5	118.6	1295.2	2.00715E+11
1983	601.0	454.8	339.0	128.9	1523.7	2.18262E+11
1984	583.4	440.7	311.7	119.6	1455.4	2.12158E+11
1985	638.3	470.5	262.0	133.6	1504.4	2.32512E+11
1986	605.6	443.2	268.3	117.1	1434.2	2.48986E+11
1987	568.6	461.7	263.6	109.6	1403.5	2.79034E+11
1988	704.9	541.1	314.7	138.5	1699.2	2.96589E+11
1989	735.7	498.5	347.6	128.6	1710.4	2.96042E+11
1990	742.9	551.4	327.0	142.6	1763.9	3.20979E+11
1991	746.8	556.9	259.9	120.2	1683.8	2.70105E+11
1992	728.6	572.1	365.9	128.2	1794.8	2.88208E+11
1993	803.0	598.4	308.2	133.0	1842.6	2.79296E+11
1994	818.1	657.7	298.8	140.4	1915.0	3.27276E+11
1995	769.8	621.0	290.3	123.1	1804.2	3.60282E+11
1996	817.3	693.5	341.1	142.4	1994.3	3.92897E+11
1997	825.4	663.5	304.0	138.3	1931.2	4.15868E+11
1998	860.8	712.9	313.3	149.1	2036.1	4.21351E+11
1999	896.8	763.7	303.4	134.1	2098.0	4.5882E+11
2000	849.8	696.8	310.8	110.7	1968.1	4.68395E+11
2001	933.4	727.7	333.7	133.7	2128.5	4.85441E+11
2002	718.2	657.6	260.7	111.3	1747.8	5.14938E+11
2003	885.3	721.6	376.0	149.1	2131.9	6.07699E+11
2004	831.3	686.4	334.6	131.3	1983.6	7.09149E+11
2005	917.9	693.5	340.7	133.8	2086.0	8.20382E+11
2006	933.6	758.1	339.2	142.0	2172.8	9.4026E+11
2007	966.9	785.7	407.5	147.6	2307.8	1.21674E+12
2008	991.8	806.8	400.4	145.7	2344.7	1.1989E+12
2009	890.9	808.0	335.5	146.6	2181.1	1.34189E+12
2010	959.8	868.7	434.0	182.4	2444.9	1.67562E+12
2011	1053.0	948.8	420.1	170.9	2592.9	1.82305E+12
2012	1052.4	935.1	400.4	183.4	2571.3	1.82764E+12
2013	1066.5	958.5	432.9	192.5	2650.4	1.85672E+12
2014	1054.8	865.3	428.6	171.5	2520.2	2.03913E+12
2015	1044.1	922.9	385.2	163.5	2515.7	2.10359E+12
2016	1097.0	985.1	437.7	231.3	2751.1	2.2948E+12
2017	1127.6	998.7	469.7	254.2	2850.1	2.65147E+12
2018	1164.8	1036.0	430.6	220.8	2852.1	2.70111E+12
2019	1188.7	1078.6	477.5	230.3	2975.0	2.8705E+12

2019	1188.7	1078.6	477.5	230.3	2975.0	2.8705E+12
2020	1222.7	1095.2	511.5	257.2	3086.5	2.66025E+12
: Data for 2020-21 are based on Fourth Advance Estimates.						
see Notes on Tables.						
ce : Ministry of Agriculture & Farmers Welfare, Government of India.						

• Methodology

Comparison by Diagrammatic Representation

Diagrams like graphs, charts, maps, pictures etc. are attractive and effective means for presentation of statistical data. It is more effective than tabular representation, being easily intelligible to a layman. Indeed, diagrams are almost essential whenever it is required to convey any statistical information to the general public. Diagrams are readily capable of revealing some features of the exhibited data.

Divided Bar Diagram

In some situations, the value of a variable are available for a number of components and comparison among different components or the relation between each part and the whole may be necessary. In this context, the proportions or percentages of the variable components are given more importance than the absolute values, proportions or percentages are expected to give a better idea of the relative importance of each components. For this purpose, one may draw a divided bar diagram. A bar of suitable length and width is taken, its total area being regarded as 100. If a vertical bar is chosen, then this area is divided into a number of sections by drawing lines parallel to the base, in such a way that the area of each sections represent the percentage for the corresponding category. By using divided bar diagram we can compare the relative contributions or importance of different categories to the study variable.

Column diagram

A column chart is a data visualization where each category is represented by a rectangle, with the height of the rectangle being proportional to the values being plotted. Column charts are also known as vertical bar charts. These have more general applicability than line diagrams in the sense that they may be used for series varying either over time or over space. In this method bars of equal width are taken for the different items of the series, drawn over base line. The length or height of a bar representing the value of the variable concerned. It is preferable to take the bars horizontally for data varying over space and vertically in the case of a series varying over time. We can compare the different items of the series by visualizing bar diagram.

Scatter diagram

Suppose that we have n pairs of values (xi,yi), i =1,2,...n of two variables X and Y. Considering two perpendicular axes of co-ordinates ,one for X and the other for Y, each pair of values is plotted as a point on a graph paper and we get a set of points . This diagram is called scatter diagram or scatter plot. Therefore scatter diagram is the simplest mode of diagrammatic representation of bivariate data. The scatter diagram serves as a useful technique in the study of the nature of relationship between the variables and for measuring the extent or degree of relationship.

Method of Time Series Analysis

Simple Moving Average

A moving average is a calculation to analyse the data points by creating a series of averages of different subsets of the full dataset. The simple moving average of period k of a time series gives us a new series of arithmetic means, each of k successive observations of the time series. We start with the first k observations. At the next stage, we leave the first and include the (k+1) st observation. This process is repeated until we arrive at the last k observations. Each of these means is centred against the time which is the mid-point of the time interval included in the calculation of the moving average. Thus when k, the period of the moving average, is odd, the moving average values correspond to tabulated time values for which the time series is given. When the period is even, the moving average fails midway between two tabulated values. In this case, we calculate a subsequent two-item moving average to make the resulting moving average values correspond to the tabulated time periods. Moving averages are an important analytical tool used to identify current trends and the potential for a change in an established trend.

Mathematical Curve Fitting

It is an essential part of the concept of trend that the movement over fairly long periods is smooth. This is perhaps the best and most rational method of determining the trend. In this case, a suitable trend equation is selected and then the constants involved in the equation are estimated on the basis of the data in hand. After derived the estimated trend equation we can predict the future trend values.

Polynomial Trend Line

Here we assume that our suitable trend equation is a polynomial in time element 't'. So here we consider a mathematical model:-

$Y_t = a_0 + a_1t + a_2t^2 + a_3t^3 + \ldots.. + a_pt^p + e_t$

where a_i's, i = 1(1)p are constants and e_t denotes a random error with $E(e_t) = 0$and $V(e_t) = \sigma^2$

Now we can estimate the constants a_i's by method of least squares and fit thepolynomial model. In this method the constants are determined by minimizing,

$$S = \sum_t \left(y_t - a_0 - a_1 t - a_2 t^2 - \cdots - a_p t^p\right)^2$$

The normal equations are,

$$\frac{\partial S}{\partial a_j} = 0, \forall\, j = 0(1)$$

$$\Rightarrow \sum_t t^j y_t = a_0 \sum_t t^j + a_1 \sum_t t^{j+1} + a_2 \sum_t t^{j+2} + \cdots + a_p \sum_t t^{j+} , j = 0(1)p$$

By solving the normal equations we can obtain the estimates of the constants a_0, a_1, a_2,..., a_p, and fit a trend equation on the time element 't

Growth Curves

The family of curves (polynomials) described above represents a simple and very useful type but a curve of $Y_t = a_0 + a_1t$ or $Yt = a_0 + a_1t + a_2t^2$ etc may not be a satisfactory description of the trend of some time series for the period shown or for the prediction purpose also. Perhaps of even greater general utility, in the analysis of time series, are curves of a semi-logarithmic (exponential) type.

Exponential Curve

The simplest exponential curve may be written as $Y_t = ab^t$, where a > 0 , b > 0

Now taking logarithm both side, $\log Y_t = \log a + t \times (\log b)$, which is a straight line in t

So, here we assume a mathematical model $\log Y_t = \log a + t \times (\log b) + e_t$, where et denotes a random error with $E(e_t) = 0$ and $V(e_t) = \sigma^2$

Now we can apply a least squares method to the logarithm of the original data (Y_t) to estimate a and b and fit an exponential trend equation.

Significance Test

Consider the regression model

$Y_t = a_0 + a_1 t_i + a_2 t_i^2 + a_3 t_i^3 + \ldots.. + a_p t_i^p + e_{ti}$; $i = 1(1)n$

we assume that $e_{ti} \sim$ iid $N(0, \sigma^2)$

So here are p covariates $t, t^2, t^3, \ldots, t^p$, the response variable Y_t and let there are n observations of pair (t, Y_t). Now for simplification we define the covariates as $t_1 = t, t_2 = t^2, t_3 = t^3, \ldots, t_p = t^p$; response $Y_t = y$ and error $e_t = e$. Therefore the model becomes,

$y_i = a_0 + a_1 t_{1i} + a_2 t_{2i} + a_3 t_{3i} + \ldots.. + a_p t_{pi} + e_i$; $i = 1(1)n$ where e_i iid $N(0, \sigma^2)$

Now we want to test whether the covariate has any significant effect on response variable or not. So, our testing problem will be,

$H_0 : a_j = 0 \ \forall\ j = 1,2,\ldots,p$

vs.

$H_1 : a_j \neq 0$ for at least one j

Now, the unrestricted sum square error (SSE) will be,

$$SSE = \min_{a_i} \sum_{i=1}^{n} (y_i - a_0 - a_1 t_{1i} - a_2 t_{2i} - a_3 t_{3i} + \cdots - a_p t_{pi})^2$$

For the sake of simplicity, we rewrite the model as,

$y_i = a_0 + a_1 t'_{1i} + a_2 t'_{2i} + a_3 t'_{3i} + \ldots.. + a_p t'_{pi} + e_i$

where, $t'_{ji} = t_{ji} - \bar{t}_j \ \forall\ j = 1,2,\ldots,p$

$$i.e., SSE = \sum_{i=1}^{n} (y_i - a_0 - a_1 t'_{1i} - a_2 t'_{2i} - a_3 t'_{3i} - \cdots - a_p t'_{pi})^2$$

$$\therefore \frac{\partial}{\partial a_0} \sum_{i=1}^{n} (y_i - a_0 - a_1 t'_{1i} - a_2 t'_{2i} - a_3 t'_{3i} - \cdots - a_p t'_{pi})^2 = 0$$

$$\sum_{i=1}^{n} (y_i - a_0 - a_1 t'_{1i} - a_2 t'_{2i} - a_3 t'_{3i} - \cdots - a_p t'_{pi}) = 0$$

$$\sum_{i=1}^{n} y_i = n a_0 + a_1 \sum_{i=1}^{n} t'_{1i} + a_2 \sum_{i=1}^{n} t'_{2i} + \cdots + a_p \sum_{i=1}^{n} t'_{pi}$$

$$Now, \sum_{i=1}^{n} t'_{ji} = \sum_{i=1}^{n} (t_{ji} - \bar{t}_j) = 0, \forall\ j = 1(1)p$$

$$\Rightarrow \hat{a}0 = \bar{y}$$

$$\frac{\partial}{\partial a_j}\sum_{i=1}^{n}(y_i - a_0 - a_1\acute{t}_{1i} - a_2\acute{t}_{2i} - a_3\acute{t}_{3i} - \cdots - a_p\acute{t}_{pi})^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n}(y_i - a_0 - a_1\acute{t}_{1i} + a_2\acute{t}_{2i} - a_3\acute{t}_{3i} - \cdots - a_p\acute{t}_{pi})\acute{t}_{ji} = 0$$

We can solve the above normal equations and obtain the value of $\hat{t}_1, \hat{t}_2, \hat{t}_3, \ldots, \hat{t}_p$

$$i.e.\ SSE = \sum_{i=1}^{n}(y_i - \hat{a}_0 - \hat{a}_1 t_{1i} - \hat{a}_2 t_{2i} - \hat{a}_3 t_{3i} - \cdots - \hat{a}_p t_{pi})^2 \ with\ df = n-k-1$$

$$Now, under\ H_0, a_j = 0\ \forall\ j\ and\ SSE_{H_0} = \min_{a_0}\sum^{n}(y_i - a_0)^2$$

Here, $\hat{a}_0^{H0} = \bar{y}$ [SD is the least RMSD]

$$i.e.\ SSE_{H_0} = \sum_{i=1}^{n}(y_i - \bar{y})^2\ with\ df = n-1$$

$$\therefore \frac{SSE_{H_0} - SSE}{\sigma^2} \sim \chi^2_{n-1-(n-k-1)} \equiv \chi^2_k\ and\ \frac{SSE}{\sigma^2} \sim \chi^2_{n-k-1}$$

Also it can be shown that SSE_{H0} – SSE is independent of SSE

$$F = \frac{\frac{SSE_{H_0} - SSE}{\sigma^2} \Big/ k}{\frac{SSE}{\sigma^2} \Big/ (n-k-1)} \sim F_{k,n-k-1}$$

So, we reject H_0 at level α if $F_{obs} > F_{\alpha;k,n-k-1}$

In case of p-value if our p-value i.e. $P(F>F_{obs}) < \alpha$ we reject our null hypothesis i.e. $a_j = 0 \ \forall \ j = 1,2,\ldots,p$.
So if our p-value < α we can say the covariates have significant effect on responseand in that case our model is statistically significant.

R^2 – Coefficient of Determination and Adjusted R Square

In statistics the coefficient of determination denoted by R^2 or r^2 and pronounced "R squared", is the proportion of the variation in the dependent variable that is predicted from the independent variable(s).

Definition:-
A data set has n values marked $y_1, y_2,\ldots y_n$(denoted as y_i) each associated with afitted(or modelled, or predicted) value $f_1,f_2,\ldots f_n$(denoted as f_i or sometimes $\hat{y}_i$)

Define the residuals as $e_i = y_i - f_i$

If $\bar{y}$ is the mean of the observed data:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Then the variability of the dataset can be measured with two sum of squaresformulas:

The total sum of squares (proportional to variance of the data):

$$SS_{to} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

The sum of squares of residuals, also called residual sum of squares:

$$SS_{res} = \sum_{i=1}^{n}(y_i - f_i)^2 = \sum_{i=1}^{n} e_i^2$$

Then the most general definition of the coefficient of determination is

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

R^2 is a statistic that will give some information about the goodness of fit of a model. In regression it measures how well the regression predictions approximate the real data points. The value of R^2 lies between 0 and 1. More the value of R^2 close to1 indicates a better fit of model. R^2 of 1 indicates that the regression predictions perfectly fit the data.

Adjusted R^2

However each time we add a new predictor variable to the model the R- squared is guarantee to increase even if the predictor variable isn't useful.

The adjusted R-squared or adjusted R^2 is a modified version of R^2 that adjusts for the number of predictors in a regression model. It is calculated as

$$1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

R^2 : The coefficient of determination of the
modeln : The number of observations
k : The number of predicted variables

Since R^2 always increases as you add more predictors to a model, adjusted R^2 can serve as a metric that tells you how useful a model is, adjusted for the number of predictors in a model. Therefore the adjusted R^2 tells us the percentage of variation explained by only the independent variables that actually affect the dependent variable. Same as R^2, the value of adjusted R^2 lies between 0 and 1. Morethe value of adjusted R^2 close to 1 indicates that all the predictor variables in modelhave better significant effects.

For the second part our analysis regarding total production (based on four food-grains) and GDP-

Objectives of the Study

The main objectives of the study are :

***To examine how agricultural production contributes to the economic growth in India (1971 - 2020).**

*** To fit the linear regression growth model to find out the relationship between gross domestic product and the selected variables of agricultural production (four major food-grains) in India.**

Data and Method of Analysis for fitting linear regression model

The present study makes use of secondary data drawn from Economic Survey of India, World Bank reports, and RBI's *Handbook of Statistics* for the period of 1971 - 2020. The linear growth regression model is employed where GDP is a dependent variable and the explanatory variables are the major crops and factors which effect the economic growth of the country. Four major crops which are included in the model are wheat, rice, pulses , coarse cereals. The data analysis is done using R software.

(1) Model Specification : The study uses linear growth regression model. The gross domestic product (GDP) is a dependent variable and the four independent variables are the major crops such as wheat, rice, pulses, coarse cereals.
The model of linear regression is as follows :

GDP = á + â$_1$ *Wheat* + â$_2$ *Rice* + â$_3$*Pulses* + â$_4$*cereals*+ u_t , , ,

where, *GDP* = is the gross domestic product and á = is a constant value, and â , â , â and â are parameters to $_{1\,2\,3\,4}$ be estimated. All explanatory variable : wheat, rice, pulses, cereals are the agricultural outputs, respectively.

(2) Variables Used in the Model :

(i) Dependent Variable - Gross Domestic Product : In a mixed economy, net foreign demand for domestic output is an important source of final spending. In the present study, the GDP at factor cost is taken as a proxy for economic growth which measures the total output in an economy. It measures how big the economy is and has been chosen as an important indicator in this case because it captures all the variables that concern economic performance. GDP measures the total production outputted in the selected period at prices of the same base year.

(ii) Independent Variables : The study analyzes the impact of all four explanatory variables on gross development product of India from 1971 –2020 rice, wheat, pulses and coarse cereals.

ADF TEST

In statistics and econometrics, an augmented Dickey–Fuller test (ADF) tests the null hypothesis that a unit root is present in a time series sample. The alternative hypothesis is different depending on which version of the test is used, but is usually stationarity or trend-stationarity. It is an augmented version of the Dickey–Fuller test for a larger and more complicated set of time series models.

The augmented Dickey–Fuller (ADF) statistic, used in the test, is a negative number. The more negative it is, the stronger the rejection of the hypothesis that there is a unit root at some level of confidence.

Time Series Analysis and Results

First we intend to analyse the annual production of rice, wheat , pulses and coarse cereals in India for past 50 years by using appropriate time series analysis method.

Studying and Comparing Data Using their Graphical Representation:

For studying and comparing the production of different food-grains over different years we use column charts for each of the food-grains and divided bar diagram .

Then we use a scatter diagram of total production and GDP to measure the correlation between them if exists.

In fig 1.1 we can see that rice production has been overall increasing over the time period 1971-2020 . We can also observe that in years 2002,2004,2009 the rice production has reduced comparing to the adjacent years . The reasons for low yielding rice in India-i. shortage of water is the major cause of low productivity of rice in India. The irrigation water supplied to the crops does not meet the needs for optimum yield. ii. monsoon is irregular and sometimes there is not enough rainfall. iii. Poor irrigation iv. Quality of soil varies etc.

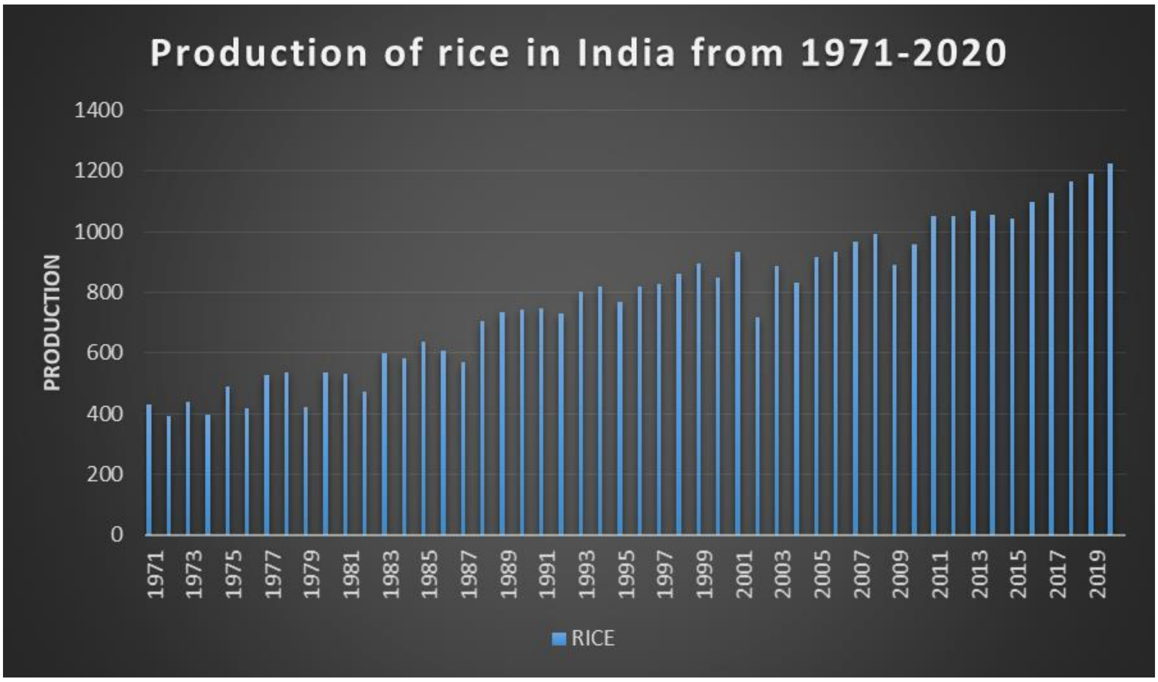


Fig: 1.1

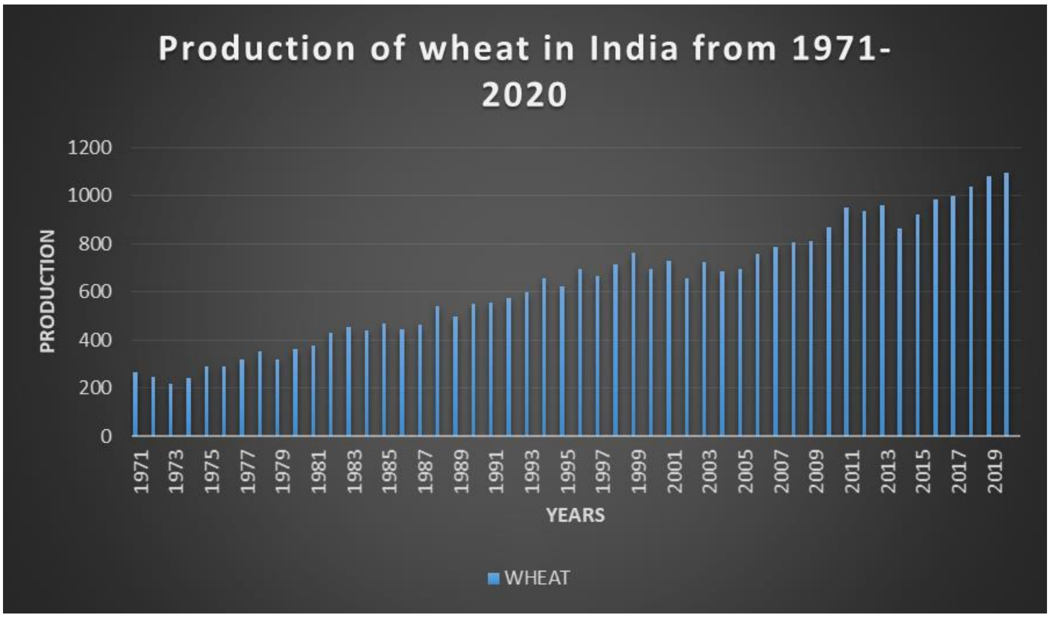


Fig: 1.2

In fig 1.2 we can see the wheat production's movement over time 1971-2020. It shows overall increasing production of wheat in India .

From fig 1.1 and 1.2 we can observe decadal changes in rice and wheat production in India. Year 2002-03, 2004-05 and 2009-10 are drought (failed monsoon) .

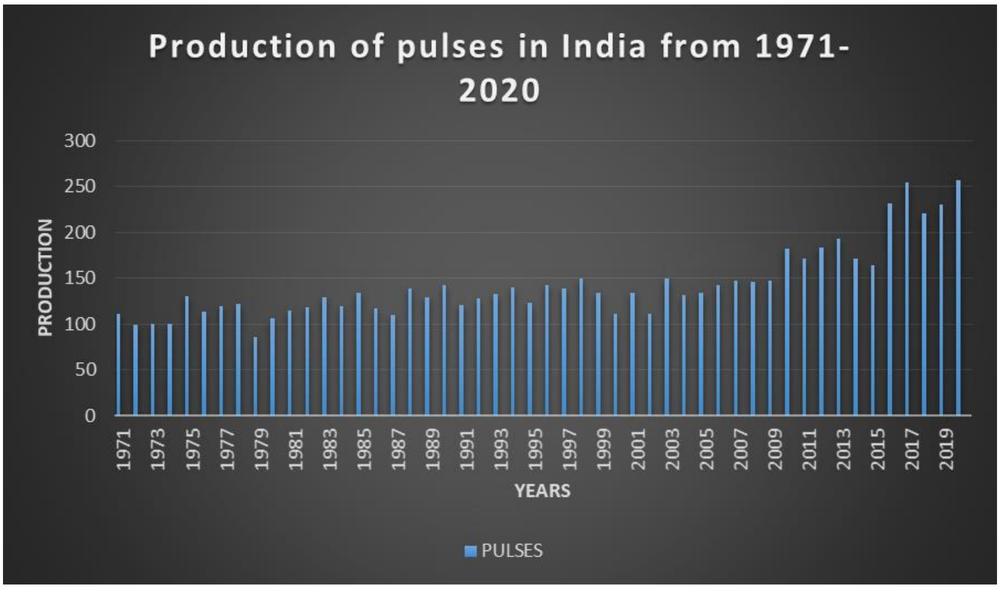


Fig: 1.3

From fig 1.3 it is clear that the pulses production has been increasing over years. As per data, During 2005-06, the total production of pulses in India was 13.38 million MT, which increased to 25.58 million MT during 2020-21. This shows an impressive growth of 91% or a compound annual growth rate (CAGR) of 4.42%. But compared to fig 1.1 and 1.2 we can see that the pulse production is much less than that of rice and wheat production in India due to many reasons. The production of pulses, however, is concentrated in a few states in India. The states of Madhya Pradesh, Rajasthan, Maharashtra, Uttar Pradesh and Karnataka account for more than 70 per cent of the total area as well as production of pulses .

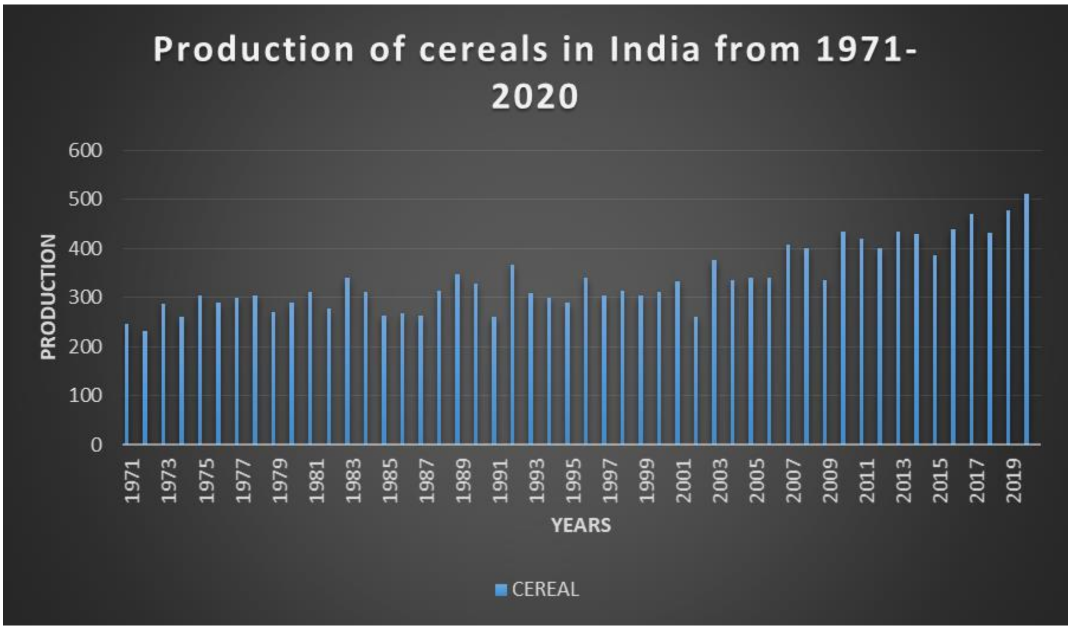


Fig : 1.4

From fig 1.4 it is evident that the production of coarse cereals is increasing over years overall.

India's agricultural output declined during the year 2002-03 to 183 million tonne compared to 212 million tonne produced in previous year, a drop of 13.9 per cent, according to the Centre for Monitoring Indian Economy. The drought has affected bajra, small millet, ragi and maize crop and production of cereals is placed at 171.71 million tonne, 14 per cent lower than 198.8 million tonne recorded in 2001-02.

In fig : 1.5, we can see that throughout the 50 years period 1971-2020 the relative contribution of four major food-grains of India has remained almost constant , has not been changed significantly .

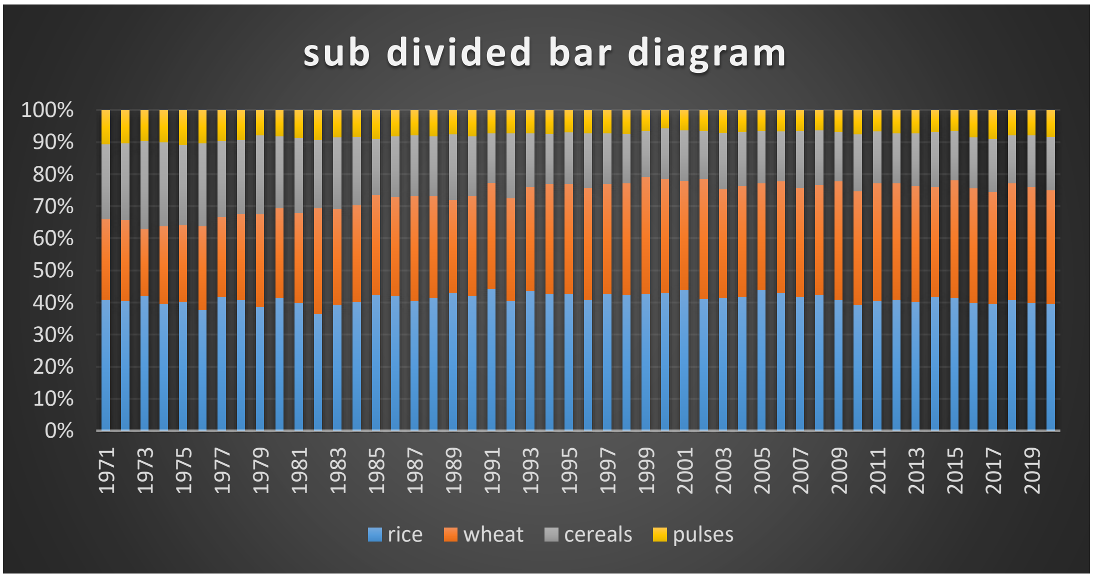


Fig: 1.5

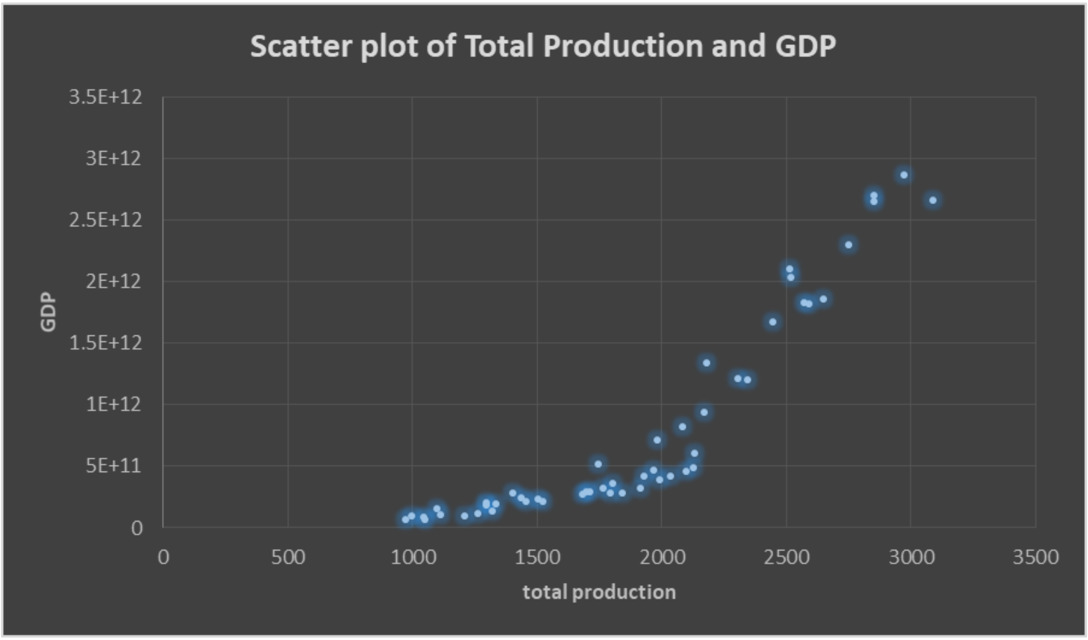


Fig : 1.6

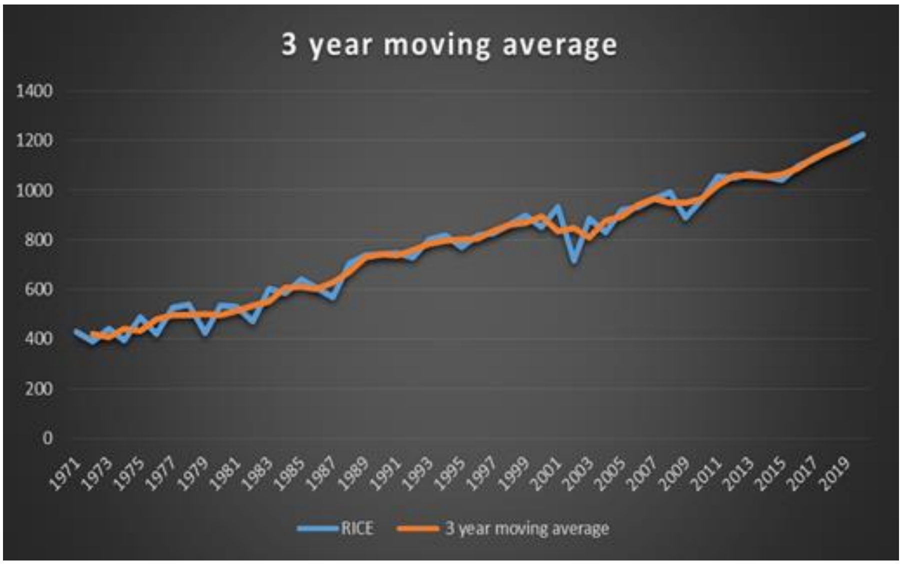
From the fig:1.6 of scatter diagram of total production in India and GDP it is evident that these two variables are positively correlated with each other over time.

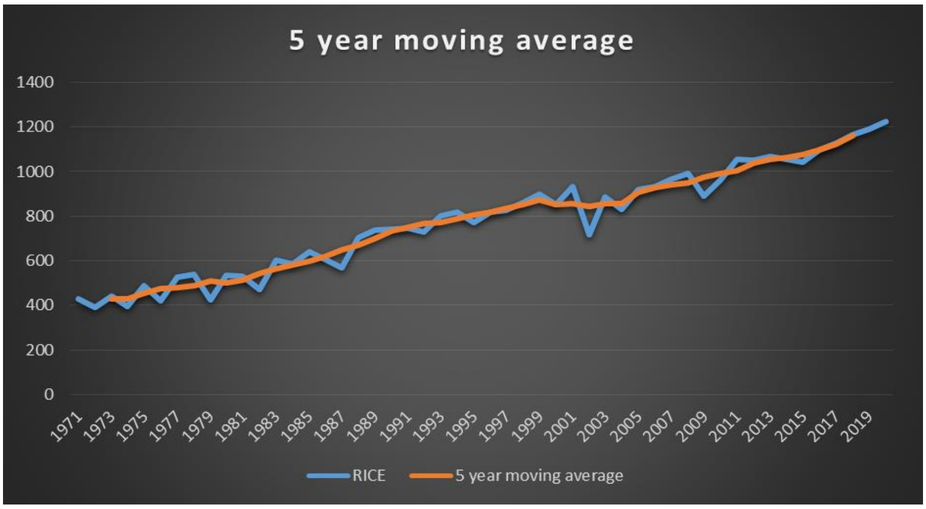
Analysis Trend by Simple Moving Average Method

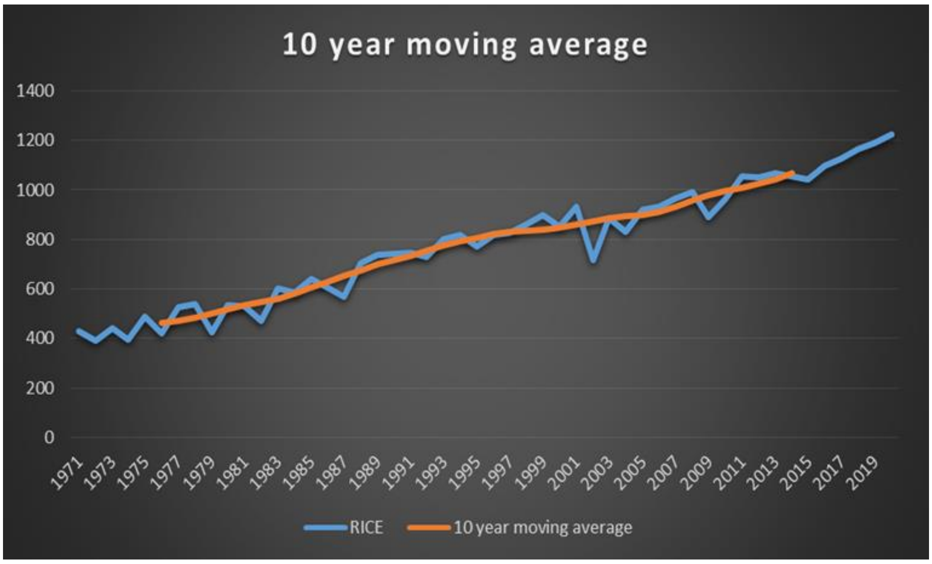
At first we use Simple Moving Average Method for analysing the trend of the individual annual production of rice , wheat, pulses and coarse cereals in India . We obtain simple moving average for 3-year, 5-year and 10-year period.

YEAR	RICE	3 year moving average	5 year moving average	calculat	10 year moving average
1971	430.7				
1972	392.4	421.2			
1973	440.5	409.5666667	429.36		
1974	395.8	441.2333333	427.06		
1975	487.4	434.1333333	453.92		
1976	419.2	477.7666667	473.36	459	464.09
1977	526.7	494.5333333	478.86	469.18	473.12
1978	537.7	495.9	488.64	477.06	485.085
1979	423.3	499.1	511.3	493.11	502.49
1980	536.3	497.3666667	500.2	511.87	519.415
1981	532.5	513.3333333	512.86	526.96	536.28
1982	471.2	534.9	544.88	545.6	547.695
1983	601	551.8666667	565.28	549.79	558.15
1984	583.4	607.5666667	579.9	566.51	582.13
1985	638.3	609.1	599.38	597.75	608.08
1986	605.6	604.1666667	620.16	618.41	629.125
1987	568.6	626.3666667	650.62	639.84	652.71
1988	704.9	669.7333333	671.54	665.58	675.68
1989	735.7	727.8333333	699.78	685.78	697.515
1990	742.9	741.8	731.78	709.25	715.825
1991	746.8	739.4333333	751.4	722.4	732.985
1992	728.6	759.4666667	767.88	743.57	756.41
1993	803	783.2333333	773.26	769.25	777.045
1994	818.1	796.9666667	787.36	784.84	792.895
1995	769.8	801.7333333	806.72	800.95	806.295
1996	817.3	804.1666667	818.28	811.64	820.97
1997	825.4	834.5	834.02	830.3	829.78
1998	860.8	861	850.02	829.26	833.375
1999	896.8	869.1333333	873.24	837.49	838.15
2000	849.8	893.3333333	851.8	838.81	846.215
2001	933.4	833.8	856.7	853.62	859.435
2002	718.2	845.6333333	843.6	865.25	872.325
2003	885.3	811.6	857.22	879.4	885.95
2004	831.3	878.1666667	857.26	892.5	892.205
2005	917.9	894.2666667	907	891.91	897.41
2006	933.6	939.4666667	928.3	902.91	908.89
2007	966.9	964.1	940.22	914.87	931.58
2008	991.8	949.8666667	948.6	948.29	957.35
2009	890.9	947.5	972.48	966.41	977.585
2010	959.8	967.9	989.58	988.76	995.07
2011	1053	1021.733333	1004.52	1001.38	1009.55
2012	1052.4	1057.3	1037.3	1017.72	1025.755
2013	1066.5	1057.9	1054.16	1033.79	1042.44
2014	1054.8	1055.133333	1062.96	1051.09	1065.98
2015	1044.1	1065.3	1078	1080.87	
2016	1097	1089.566667	1097.66		
2017	1127.6	1129.8	1124.44		
2018	1164.8	1160.366667	1160.16		
2019	1188.7	1192.066667			
2020	1222.7				

Fig 2 : Simple Moving Average of 3 Year, 5 Year and 10 Year of Annual Rice Production

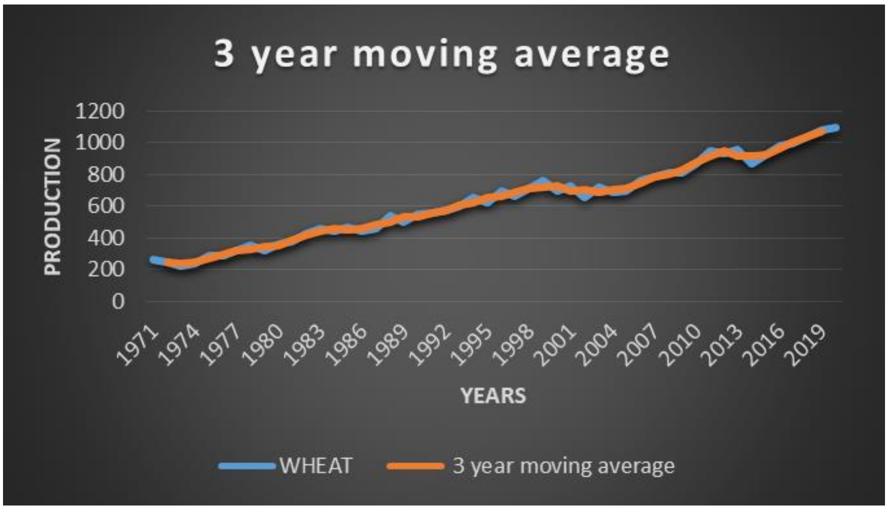


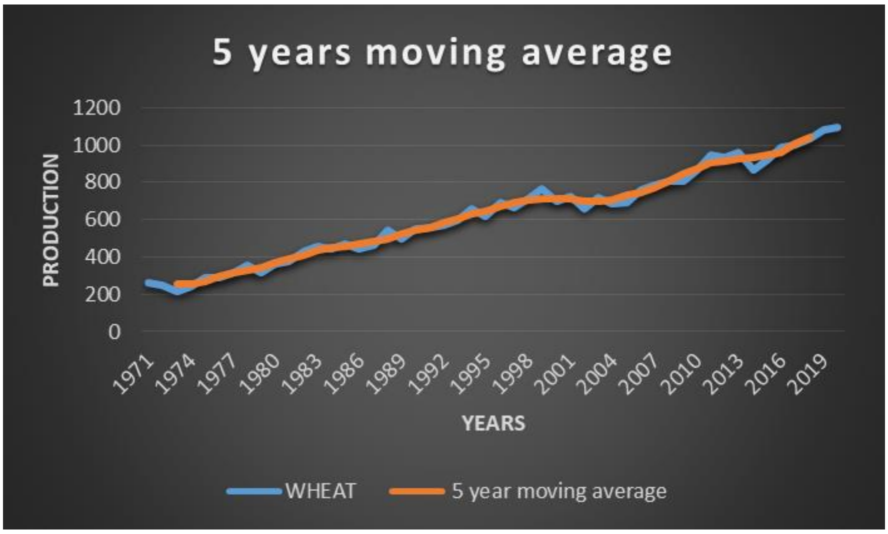


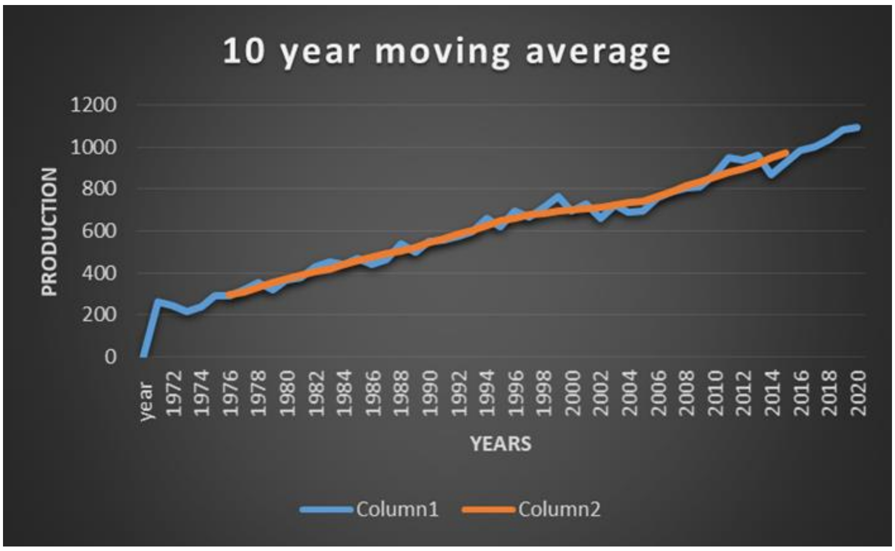


year	WHEAT	3 year moving average	5 year moving average	calculation	10 year moving average
1971	264.1				
1972	247.4	243.1			
1973	217.8	235.4	251.74		
1974	241	249.0666667	256.94		
1975	288.4	273.1666667	270.96	290.28	
1976	290.1	298.6666667	298.42	301.32	295.8
1977	317.5	320.9	313.88	319.37	310.345
1978	355.1	330.3	328.82	343.07	331.22
1979	318.3	345.5	345.7	363.04	353.055
1980	363.1	351.9666667	367.78	381.25	372.145
1981	374.5	388.5	387.72	396.56	388.905
1982	427.9	419.0666667	412.2	410.98	403.77
1983	454.8	441.1333333	433.68	429.58	420.28
1984	440.7	455.3333333	447.42	447.6	438.59
1985	470.5	451.4666667	454.18	466.43	457.015
1986	443.2	458.4666667	471.44	484.67	475.55
1987	461.7	482	483	499.09	491.88
1988	541.1	500.4333333	499.18	513.45	506.27
1989	498.5	530.3333333	521.92	535.15	524.3
1990	551.4	535.6	544	550.2	542.675
1991	556.9	560.1333333	555.46	575.23	562.715
1992	572.1	575.8	587.3	595.41	585.32
1993	598.4	609.4	601.22	612.59	604
1994	657.7	625.7	628.54	639.11	625.85
1995	621	657.4	646.82	653.65	646.38
1996	693.5	659.3333333	669.72	670.73	662.19
1997	663.5	689.9666667	690.92	679.28	675.005
1998	712.9	713.3666667	706.08	691.6	685.44
1999	763.7	724.4666667	712.92	694.47	693.035
2000	696.8	729.4	711.74	701.72	698.095
2001	727.7	694.0333333	713.48	708.18	704.95
2002	657.6	702.3	698.02	720.4	714.29
2003	721.6	688.5333333	697.36	729.79	725.095
2004	686.4	700.5	703.44	734.22	732.005
2005	693.5	712.6666667	729.06	751.41	742.815
2006	758.1	745.7666667	746.1	773.52	762.465
2007	785.7	783.5333333	770.42	801.27	787.395
2008	806.8	800.1666667	805.46	824.96	813.115
2009	808	827.8333333	843.6	842.85	833.905
2010	868.7	875.1666667	873.48	865.79	854.32
2011	948.8	917.5333333	903.82	888.49	877.14
2012	935.1	947.4666667	915.28	909.79	899.14
2013	958.5	919.6333333	926.12	932.71	921.25
2014	865.3	915.5666667	933.38	959.77	946.24
2015	922.9	924.4333333	946.1	982.42	971.095
2016	985.1	968.9	961.6		
2017	998.7	1006.6	1004.26		
2018	1036	1037.766667	1038.72		
2019	1078.6	1069.933333			
2020	1095.2				

Fig 3: Simple Moving Average of 3Year, 5 Year and 10 Year of Wheat Production in India

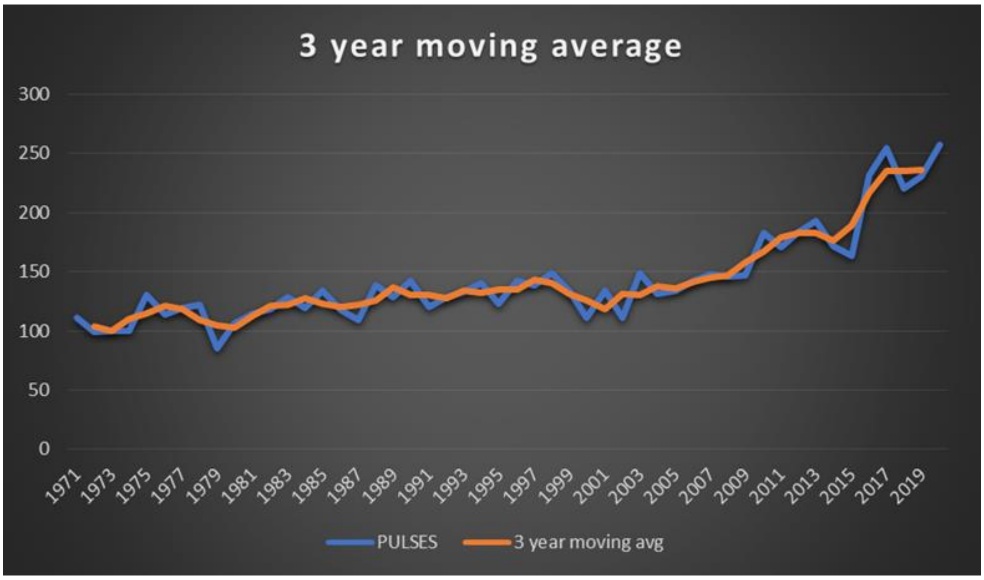


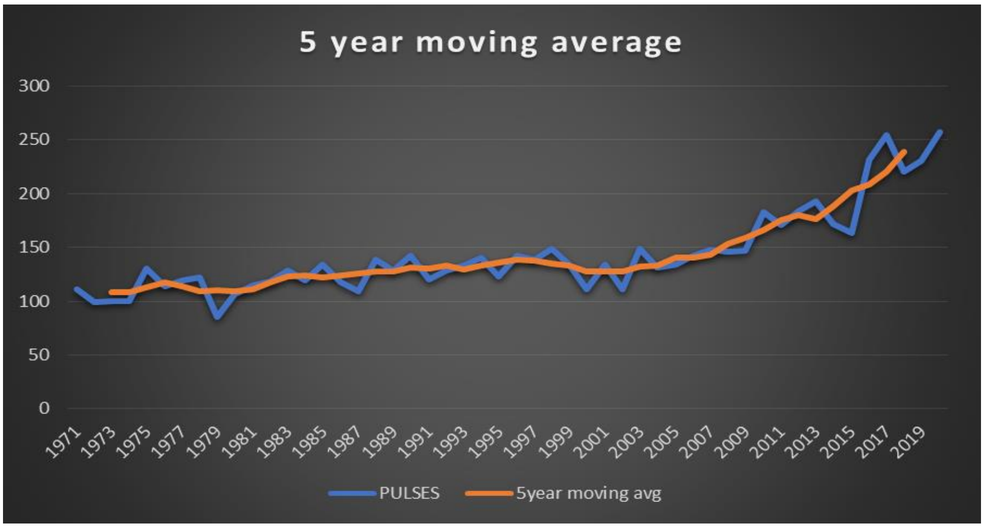


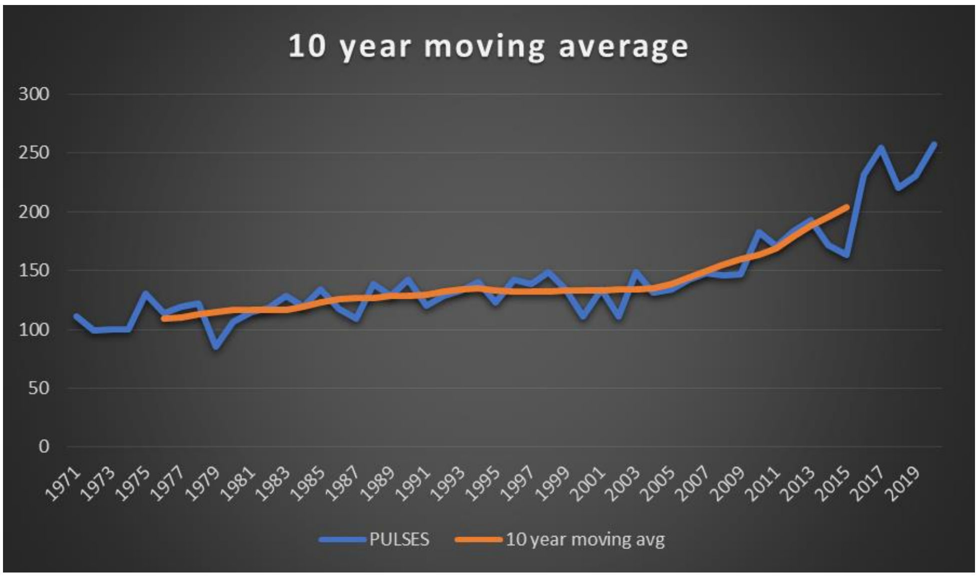


YEAR	PULSES	3 year moving avg	5year moving avg	calculation	10 year moving avg
1971	110.9				
1972	99.1	103.3666667			
1973	100.1	99.8	108.14		
1974	100.2	110.2333333	108.68		
1975	130.4	114.7333333	112.8	108.78	
1976	113.6	121.2333333	117.14	109.2	108.99
1977	119.7	118.3666667	114.24	111.15	110.175
1978	121.8	109.0666667	109.42	114.03	112.59
1979	85.7	104.6	109.72	115.97	115
1980	106.3	102.3666667	109.5	116.29	116.13
1981	115.1	113.3333333	110.92	116.64	116.465
1982	118.6	120.8666667	117.7	115.63	116.135
1983	128.9	122.3666667	123.16	117.3	116.465
1984	119.6	127.3666667	123.56	121.59	119.445
1985	133.6	123.4333333	121.76	125.22	123.405
1986	117.1	120.1	123.68	125.73	125.475
1987	109.6	121.7333333	125.48	126.69	126.21
1988	138.5	125.5666667	127.28	127.1	126.895
1989	128.6	136.5666667	127.9	129.18	128.14
1990	142.6	130.4666667	131.62	128.13	128.655
1991	120.2	130.3333333	130.52	130.66	129.395
1992	128.2	127.1333333	132.88	133.53	132.095
1993	133	133.8666667	128.98	134.59	134.06
1994	140.4	132.1666667	133.42	135.14	134.865
1995	123.1	135.3	135.44	131.95	133.545
1996	142.4	134.6	138.66	133.3	132.625
1997	138.3	143.2666667	137.4	131.61	132.455
1998	149.1	140.5	134.92	133.22	132.415
1999	134.1	131.3	133.18	132.31	132.765
2000	110.7	126.1666667	127.78	133.38	132.845
2001	133.7	118.5666667	127.78	133.34	133.36
2002	111.3	131.3666667	127.22	134.27	133.805
2003	149.1	130.5666667	131.84	133.93	134.1
2004	131.3	138.0666667	133.5	135.18	134.555
2005	133.8	135.7	140.76	142.35	138.765
2006	142	141.1333333	140.08	146.07	144.21
2007	147.6	145.1	143.14	153.28	149.675
2008	145.7	146.6333333	152.86	157.62	155.45
2009	146.6	158.2333333	158.64	161.64	159.63
2010	182.4	166.6333333	165.8	164.61	163.125
2011	170.9	178.9	175.16	173.54	169.075
2012	183.4	182.2666667	180.14	184.2	178.87
2013	192.5	182.4666667	176.36	191.71	187.955
2014	171.5	175.8333333	188.44	200.08	195.895
2015	163.5	188.7666667	202.6	207.56	203.82
2016	231.3	216.3333333	208.26		
2017	254.2	235.4333333	220.02		
2018	220.8	235.1	238.76		
2019	230.3	236.1			
2020	257.2				

Fig 4: Simple Moving average of 3Year, 5Year and 10Year of Pulses Production in India

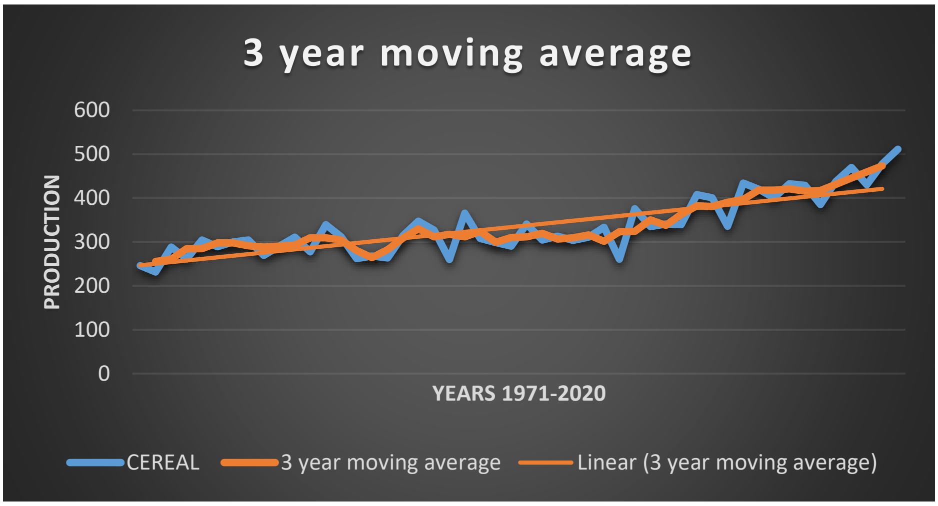


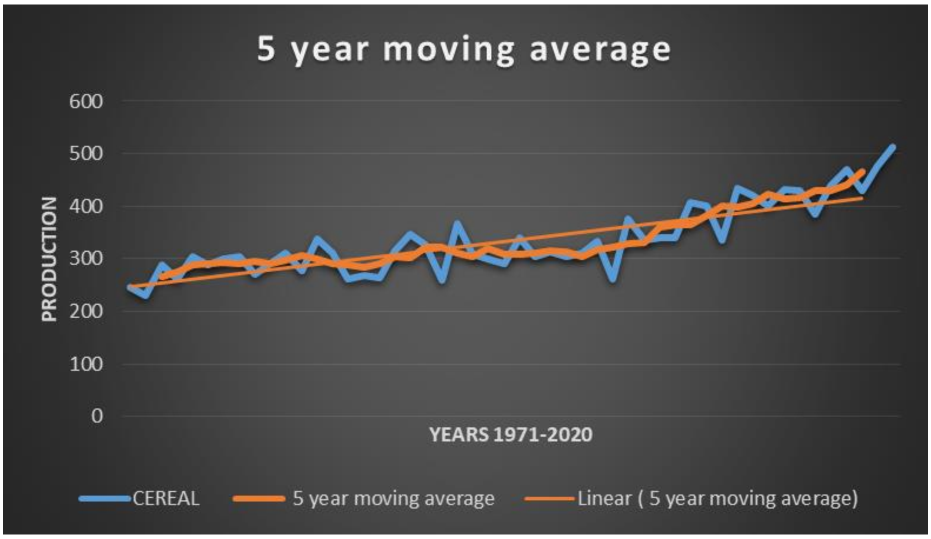


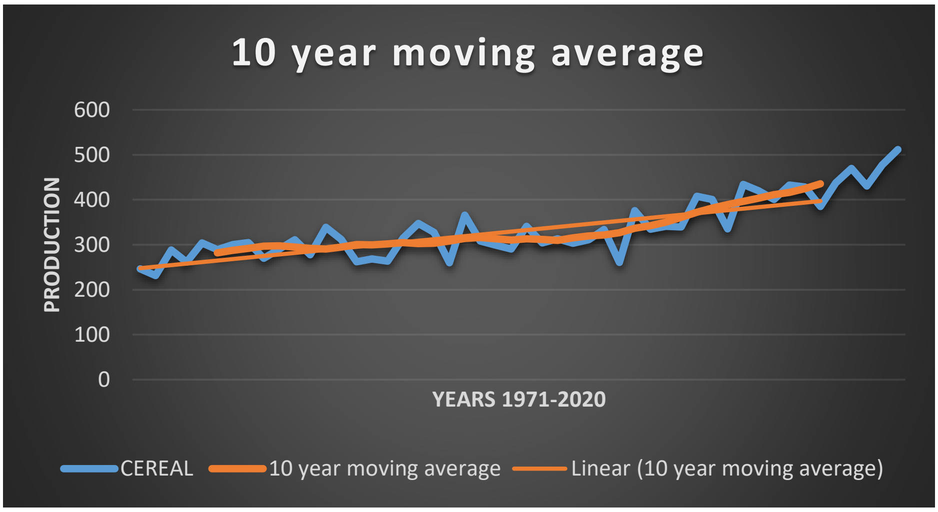


YEAR	coarse c	3 year moving average	5 year moving average	calculat	10 year moving average
1971	246				
1972	231.4	255.2333333			
1973	288.3	260.3333333	266.22		
1974	261.3	284.5666667	274.78	278.44	
1975	304.1	284.7333333	288.54	284.93	
1976	288.8	297.7	291.76	289.54	281.685
1977	300.2	297.8	293.44	294.61	287.235
1978	304.4	291.4333333	290.66	299.65	292.075
1979	269.7	288.1	295.08	295.44	297.13
1980	290.2	290.2666667	290.54	293.39	297.545
1981	310.9	292.8666667	297.46	289.73	294.415
1982	277.5	309.1333333	305.86	290.76	291.56
1983	339	309.4	300.22	298.55	290.245
1984	311.7	304.2333333	291.7	302.23	294.655
1985	262	280.6666667	288.92	297.13	300.39
1986	268.3	264.6333333	284.06	305.97	299.68
1987	263.6	282.2	291.24	302.89	301.55
1988	314.7	308.6333333	304.24	301.6	304.43
1989	347.6	329.7666667	302.56	304.43	302.245
1990	327	311.5	323.02	311.71	303.015
1991	259.9	317.6	321.72	315.75	308.07
1992	365.9	311.3333333	311.96	315.61	313.73
1993	308.2	324.3	304.62	311.19	315.68
1994	298.8	299.1	320.86	309.57	313.4
1995	290.3	310.0666667	308.48	316.95	310.38
1996	341.1	311.8	309.5	306.43	313.26
1997	304	319.4666667	310.42	313.21	311.69
1998	313.3	306.9	314.52	316.79	309.82
1999	303.4	309.1666667	313.04	321.83	315
2000	310.8	315.9666667	304.38	321.64	319.31
2001	333.7	301.7333333	316.92	331.99	321.735
2002	260.7	323.4666667	323.16	340.7	326.815
2003	376	323.7666667	329.14	343.91	336.345
2004	334.6	350.4333333	330.24	356.23	342.305
2005	340.7	338.1666667	359.6	364.87	350.07
2006	339.2	362.4666667	364.48	378.84	360.55
2007	407.5	382.3666667	364.66	384.53	371.855
2008	400.4	381.1333333	383.32	393.93	381.685
2009	335.5	389.9666667	399.5	398.38	389.23
2010	434	396.5333333	398.08	408.23	396.155
2011	420.1	418.1666667	404.58	414.45	403.305
2012	400.4	417.8	423.2	417.47	411.34
2013	432.9	420.6333333	413.44	431.67	415.96
2014	428.6	415.5666667	416.96	439.42	424.57
2015	385.2	417.1666667	430.82		435.545
2016	437.7	430.8666667	430.36		
2017	469.7	446	440.14		
2018	430.6	459.2666667	465.4		
2019	477.5	473.2			
2020	511.5				

Fig 5: Simple Moving Average of 3Year , 5 Year and 10 Year of Coarse Cereals Production in India







<u>INTERPRETATION :</u> The moving averages of 3 years, 5 years and 10 years have smoothed our data and make the trend clearer in each of the four food-grains. The longer the period is chosen the more smooth the data. So, here 10 year moving average makes the data smoothest as it takes more data points in calculation. These figures capture the natural underlying trend of our data for all the four food-grains and are giving us the general idea of the movement. The figures show us that our time series data in all of the four cases have overall an upward trend.

Fitting a mathematical curve

From the graphical visualization of our time series data for all the four food-grains and after smoothing the data by moving average we observed that the trend of our time series data for all the four food-grains are the long-term movement is entirely increasing and there is not any sign of changing its direction in downward. It is obvious that total production based on those four food-grains has also upward trend. So for polynomial trend line, we have successively fitted Linear and Quadratic model and not any polynomial higher degree than 2 for total production data. We have performed their ANOVA testing and made decision of accepting or rejecting basis on p values provided by the test. Here for the ANOVA testing is 1% level of significance is chosen. We have also observed the value of coefficient of determination i.e. the R^2 values and adjusted R^2 value in each case.

year	Yt=total	t	t^2	lnYt	T(fitted linear trend)	T'(fitted c	T"(fitted exponential trer
1971	1051.7	-25	625	6.958163	952.917276	1051.7	1051.634
1972	970.3	-24	576	6.877605	990.360905	1059.0266	1099.246554
1973	1046.7	-23	529	6.953398	1027.804534	1087.2242	1122.115101
1974	998.3	-22	484	6.906054	1065.248163	1115.8172	1145.459403
1975	1210.3	-21	441	7.098624	1102.691792	1144.8056	1169.289355
1976	1111.7	-20	400	7.013646	1140.135421	1174.1895	1193.615062
1977	1264.1	-19	361	7.142116	1177.57905	1203.9687	1218.446836
1978	1319	-18	324	7.184629	1215.022679	1234.1434	1243.795208
1979	1097	-17	289	7.000334	1252.466308	1264.7135	1269.670922
1980	1295.9	-16	256	7.166961	1289.909937	1295.6791	1296.084951
1981	1333	-15	225	7.195187	1327.353566	1327.04	1323.048493
1982	1295.2	-14	196	7.16642	1364.797195	1358.7964	1350.57298
1983	1523.7	-13	169	7.328897	1402.240824	1390.9482	1378.670082
1984	1455.4	-12	144	7.283036	1439.684452	1423.4954	1407.351711
1985	1504.4	-11	121	7.316149	1477.128081	1456.438	1436.630029
1986	1434.2	-10	100	7.268362	1514.57171	1489.7761	1466.517448
1987	1403.5	-9	81	7.246724	1552.015339	1523.5095	1497.02664
1988	1699.2	-8	64	7.437913	1589.458968	1557.6384	1528.17054
1989	1710.4	-7	49	7.444483	1626.902597	1592.1627	1559.962353
1990	1763.9	-6	36	7.475283	1664.346226	1627.0824	1592.415558
1991	1683.8	-5	25	7.428808	1701.789855	1662.3976	1625.543914
1992	1794.8	-4	16	7.492649	1739.233484	1698.1081	1659.361467
1993	1842.6	-3	9	7.518933	1776.677113	1734.2141	1693.882554
1994	1915	-2	4	7.557473	1814.120742	1770.7155	1729.121813
1995	1804.2	-1	1	7.497873	1851.564371	1807.6124	1765.094184
1996	1994.3	1	1	7.598048	1926.451629	1882.5923	1839.299585
1997	1931.2	2	4	7.565897	1963.895258	1920.6754	1877.564076
1998	2036.1	3	9	7.618791	2001.338887	1959.1539	1916.624616
1999	2098	4	16	7.64874	2038.782516	1998.0278	1956.497764
2000	1968.1	5	25	7.584824	2076.226145	2037.2971	1997.200428
2001	2128.5	6	36	7.663173	2113.669774	2076.9619	2038.749863
2002	1747.8	7	49	7.466113	2151.113403	2117.0221	2081.163685
2003	2131.9	8	64	7.664769	2188.557032	2157.4777	2124.459878
2004	1983.6	9	81	7.592669	2226.000661	2198.3287	2168.656798
2005	2086	10	100	7.643004	2263.44429	2239.5751	2213.773184
2006	2172.8	11	121	7.683772	2300.887919	2281.217	2259.828163
2007	2307.8	12	144	7.74405	2338.331548	2323.2543	2306.841263
2008	2344.7	13	169	7.759913	2375.775176	2365.687	2354.832416
2009	2181.1	14	196	7.687585	2413.218805	2408.5151	2403.821969
2010	2444.9	15	225	7.80176	2450.662434	2451.7386	2453.830693
2011	2592.9	16	256	7.860532	2488.106063	2495.3576	2504.87979
2012	2571.3	17	289	7.852167	2525.549692	2539.372	2556.990905
2013	2650.4	18	324	7.882466	2562.993321	2583.7818	2610.18613
2014	2520.2	19	361	7.832094	2600.43695	2628.587	2664.488021
2015	2515.7	20	400	7.830306	2637.880579	2673.7876	2719.919599
2016	2751.1	21	441	7.919756	2675.324208	2719.3837	2776.504367
2017	2850.1	22	484	7.955109	2712.767837	2765.3752	2834.266315
2018	2852.1	23	529	7.955811	2750.211466	2811.7621	2893.229933
2019	2975	24	576	7.997999	2787.655095	2858.5444	2953.420221
2020	3086.5	25	625	8.034793	2825.098724	2905.7221	3014.862698
total	94450.4	0	11050	374.8039			

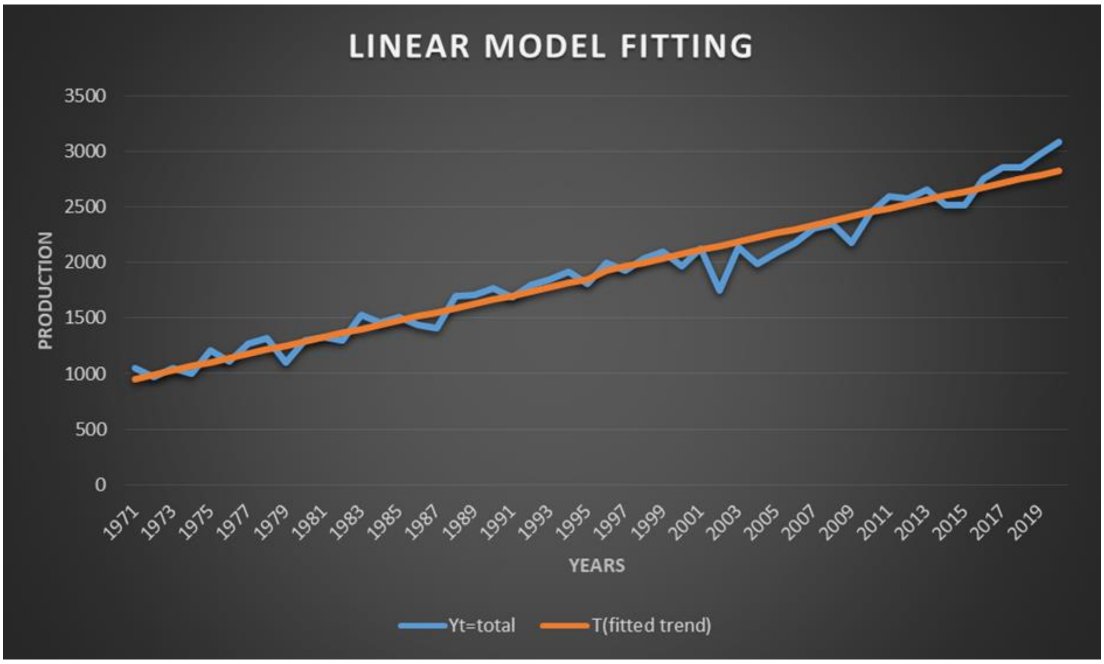
Linear Model

The fitted linear model is :

Yt= 1889.008+37.443629*t , with origin at middle of 1995 , and 1 year unit of t

Table : 2.1

SUMMARY OUTPUT	Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8
Regression Statistics								
Multiple R	0.9781687							
R Square	0.9568139							
Adjusted R Square	0.9559142							
Standard Error	120.69703							
Observations	50							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regression	1	15492380	15492380	1063.4693	2.073E-34			
Residual	48	699253.14	14567.774					
Total	49	16191633						
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1889.008	17.069138	110.66804	1.802E-59	1854.6882	1923.3278	1854.6882	1923.3278
X Variable 1	37.443629	1.1481945	32.610877	2.073E-34	35.135029	39.752229	35.135029	39.752229



Since the p-value < 0.01 ,we accept the linear model fitting.

Quadratic Model:

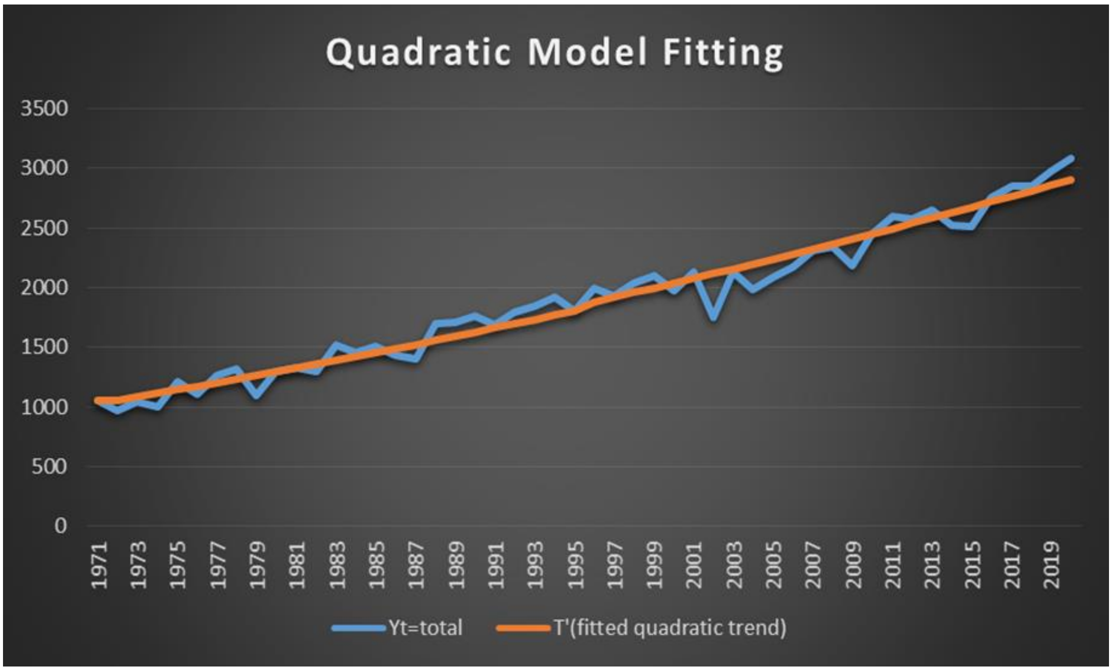
The fitted quadratic model is –

Yt= 1844.9046 +37.489954*t+0.1977099*t^2 , with origin at middle of 1995 and 1 year unit of t

Table : 2.2

SUMMARY OUTPUT	Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8
Regression Statistics								
Multiple R	0.979676							
R Square	0.9597651							
Adjusted R Square	0.9580158							
Standard Error	116.34696							
Observations	50							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	14853556	7426778.2	548.64367	8.051E-33			
Residual	47	622684.3	13536.615					
Total	49	15476241						
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1844.9046	25.253948	73.054107	3.163E-49	1794.071	1895.7382	1794.071	1895.7382
-25	37.489954	1.1436439	32.781143	1.524E-33	35.187918	39.791989	35.187918	39.791989
625	0.1977099	0.0895376	2.2081208	0.0322598	0.01748	0.3779398	0.01748	0.3779398

since the p-value is < 0.01 , we accept the quadratic model fitting.



Exponential model: For growth curve now we are fitting exponential model and performing ANOVA testing for the linear model : log Yt = log a + t × (log b) and made decision of accepting or rejecting basis on p values provided by the test. Here for the ANOVA testing is 1% level of significance is chosen.

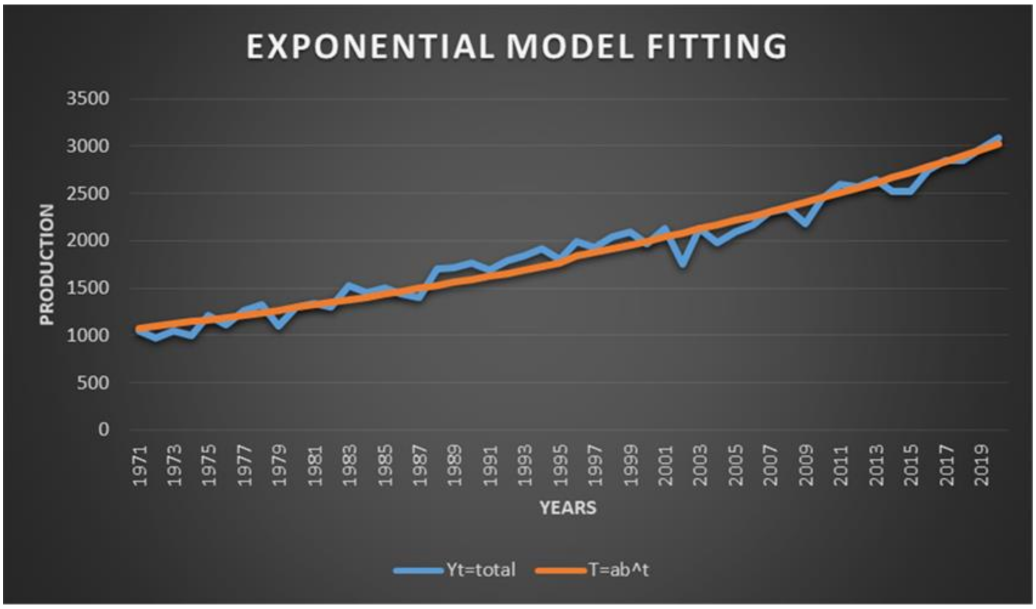
The fitted exponential model is:-

Yt = exp(0.025904)* exp(7.4965497)^t , with origin at middle of 1995 and 1 year unit of t

Table : 2.3

Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8	Column9
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.9749722							
R Square	0.9505708							
Adjusted R S	0.9495191							
Standard Err	0.0698856							
Observation:	50							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	4.4144203	4.4144203	903.85464	2.414E-32			
Residual	48	0.2295477	0.004884					
Total	49	4.643968						
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	7.4965497	0.0099898	750.4229	1.651E-97	7.4764529	7.5166465	7.4764529	7.5166465
-25	0.0205904	0.0006849	30.064175	2.414E-32	0.0192126	0.0219682	0.0192126	0.0219682

Since the p-value < 0.01 , we accept the exponential model fitting.



Now our time series data the amount of production has no upper limit. From the time series plot graph and after smoothing the data we can see the series is increasing and the amount of growth doesn't decline by a constant percentages and approaches any upper limit. That is there is not any decrease in growth rate or increasing rate. So we need not to check the fitting of Modified Exponential model and Gompertz Model.

Now from Table 2.1, Table 2.2, Table 2.3 we can see that among linear, quadratic, exponential model; quadratic model has the largest R2 value i.e. 0.9597651 and largest adjusted R2 value which is 0.9580158. So we conclude that the fitted quadratic trend line equation is best for our time series data. Now, we can predict the total production (based on four major food-grains) for future years through our fitted quadratic trend model.

Predicted Value of Total Production of India for Future Years

year	t	production(million tonnes)
2021	26	2953.295296
2022	27	3001.263875
2023	28	3049.627874
2024	29	3098.387292
2025	30	3147.54213
2026	31	3197.092388
2027	32	3247.038066
2028	33	3297.379163
2029	34	3348.11568
2030	35	3399.247618
2031	36	3450.774974

Analysis and Results for fitting linear regression model

The data analysis is done using R software. The data under the study is a time series data, therefore, first the stationarity of the data series is tested through the unit root method using Augmented Dickey - Fuller test (ADF). The ADF test is applied on all the variables under the study. The test hypothesis for the unit root ADF test is as under :

Null Hypothesis : H : ä = 0 (i.e., there is a unit root or time series is non - stationary).
Alternative Hypothesis : H : ä < 0 (the time series is stationary). 1

The test results reveal that at the I(0) levels, the absolute value of the ADF test is lower than the absolute critical value at 95% level of confidence for all the variables under the study. This gives enough ground for not rejecting the null hypothesis, meaning that the data series under the study are non - stationary. This violates the basic assumption of the time series. In time series analysis, estimating a linear equation or developing a model on non - stationary data may yield spurious results. Therefore, to check whether the model is spurious or not, the model is estimated with the original non - stationary data, assuming that the data will be stationary at the first difference level. A simple check is applied that if the R square value is greater than the Durbin - Watson value, then it may suggest that the model is spurious. If the R square value is less than the Durbin - Watson value, then the model in not spurious and can be used for further testing and analysis. The result of estimation at I(0) level indicates that the value of Durbin - Watson is 1.2371, which is higher than the R square value of 0.9066. This indicates that the model is not spurious and cannot be outrightly rejected.

The model estimation at I(0) level indicates that the model is not spurious and cannot be outrightly rejected, but the basic assumption of stationarity of data is violated, therefore, the unit root ADF test is applied at the I(1) level to check whether the data is stationary at I(1) level or not including the intercept. The summary result of the unit root ADF test applied to all the variables under the study is displayed in Table 1.

The summary of the unit root ADF test as shown in Table 1 indicates that all the variables under the study have become stationary at the I(1) order of integration at the 5% level of significance. Therefore, from the test results, it is concluded that the all the variables are stationary at I(1) level and can be used for the time series analysis and for estimating the model.

TABLE: SUMMARY OF UNIT ROOT ADF TEST RESULTS

VARIABLES	VARIABLE DESCRIPTION	ADF TEST	ORDER OF INTEGRATION	DECISION	LEVEL OF SIGNIFICANCE
GDP	Gross domestic product	-2.9107	1	stationary	5%
X1	RICE	-5.8603	1	stationary	5%
X2	WHEAT	-4.6925	1	stationary	5%
X3	COARSE CEREALS	-4.8748	1	stationary	5%
X4	PULSES	-5.2197	1	stationary	5%

However, the estimating equation results are not very promising as far as it relates to the explanatory power of the independent variables as only four out of six variables have shown a significant effect, and that too, in the short run. This may be because of the fact that the GDP is also affected by the variables other than the variables used in the present model. So, there exists scope for future research by conducting similar studies with more number of variables.

Suggestions for Policy Makers

The empirical findings of the study reveal that agricultural production has a direct effect on agricultural income and employment, and on the other hand, it also shows that economic growth has a direct relationship with agricultural production in the Indian perspective. The study demonstrates a positive relationship between the dependent variable : gross domestic product (GDP) and the selected four independent variables in the Indian economy. The study emphasizes that if there is no increase in agricultural production in India, then it has a negative impact on Indian economic growth. Inspite of the various steps taken up by the GoI from 1971 - 2020 to enhance the agricultural output in India, the agricultural sector is facing major problems in India. In order to completely remove the agricultural issues, the following can be considered in the policy making :

It is the need of hour to make an effective plan for awareness among the farmers about Zero Budget Natural Farming education. The study suggests to build effective grassroots institutions to implement the ZBNF methods all over the country. It will bring about transformation in the Indian agriculture sector.

Opportunities for informal learning need to be made more available at village level organizations (VLOs).

Agricultural production and farm income in India are frequently affected by natural disasters such as droughts, floods, cyclones, storms, cloudbursts, landslides, and earthquakes.

Susceptibility of Indian agriculture to these disasters is compounded by the outbreak of epidemics and man made disasters such as fire, sale of spurious seeds, replacing the use of fertilizers and pesticides by ZBNF adoption. It is the need of the hour to find out the feasibility of insurance adoption regarding the advantages and disadvantages of the newly introduced scheme - Pradhan Mantri Fasal Bima Yojana (previously NAIS). It is important to examine the evident implications for the choice of crops, crop productivities, variety of yields,variety of seeds and availability of HYV and GM seeds, e-NAM, APMCs for the benefit of especially the small and marginal farmers for farm credit.

A rethink on the minimum support price (MSP) process is required. It is again a major problem in India as the farmers face this condition irrespective of crop failure or bumper harvest as there has been a sharp fall in the market prices of crops like pulses, onion, potato, oilseeds, cotton, and maize. The farmers have suffered huge loss of income from 1971 - 2020 due to this.

There is a need to make proper plans for linking the rivers for resolving irrigation issues as well as to resolve the flood and drought issues in India. The study advocates for quick execution for linkage of the rivers as soon as possible.

Limitations of the Study and Scope for Further Research

The study is limited because it used secondary data drawn from various sources (as discussed) for the selected duration of the study. The data relating to agricultural production of four major crops were collected on the basis of the available sources. The study is further limited due to the selection of only four crops and ascertaining their impact on economic growth in the last more than five decades. Despite its limitations, the present study provides an opportunity to farmers, researchers, traders, and policymakers for proper impactful execution of the agricultural policies for enhancement of the agricultural output in India.

CONCLUSION :

From graphical visualization and moving average method we get a clear idea that production of each of the four major food-grains has upward trend basis on the data over the time period 1971-2020. Hence it is obvious that the total production data based on those four food-grains has also upward trend. Then we intend to fit a mathematical curve using linear model, quadratic model, exponential model on the total production data so that we can predict the production for future years. Since the value of R^2 and adjusted R^ is greatest for quadratic model , we can say that quadratic model fits the data best and based on quadratic model we predict the total production for upcoming 10 years.

Now, we all know that Gross Domestic Product (GDP) is the most commonly used measure for the size of an economy. So we want to examine how agricultural production contributes to the economic growth in India (1971 - 2020) and so fit the linear regression growth model to find out the relationship between gross domestic product and the selected variables of agricultural production (four major food-grains) in India. The study uses linear growth regression model. The gross domestic product (GDP) is a dependent variable and the four independent variables are the major crops such as wheat, rice, pulses, coarse cereals. . The data under the study is a time series data, therefore, first the stationarity of the data series is tested through the unit root method using Augmented Dickey - Fuller test (ADF). The ADF test is applied on all the variables under the study. Based on adf test on first difference level and durbin Watson value we can say that we cannot outrightly reject the null hypothesis that the time series is stationary and hence the model is not spurious .

However, the estimating equation results are not very promising as far as it relates to the explanatory power of the independent variables as only four out of six variables have shown a significant effect, and that too, in the short run. This may be because of the fact that the GDP is also affected by the variables other than the variables used in the present model. So, there exists scope for future research by conducting similar studies with more number of variables.

• Reference:-

□ Sources of Theory:-

Fundamental of Statistics (Volume One & Two) – A.M. Goon, M.K. Gupta, B. Dasgupta

Introduction to Statistics – Prasanta Kumar Giri, Jiban Banerjee https://www.statisticshowto.com/adjusted-r2/

https://www.statology.org/multiple-r-vs-r-squared/

https://outlier.ai/trendlines-moving-averages/

□ Software Used:-

Microsoft Excel 2013

R Statistical Software

Microsoft Word 2013

□ Sources of Data:-

https://rbi.org.in/

• Acknowledgement:-

I am indebted to so many people for helping me in the preparation of this project.

Firstly, Dr. Apurba Roy, Vice- Principal , Asutosh College, University of Calcutta, without whose help I couldn't have been a part of this prestigious college.

I owe a deep debt of gratitude to my supervisor Oindrila Bose for necessary guidance, for this presentation of this dissertation, valuable comments and suggestions. I am extremely grateful to him for the necessary stimulus, support and valuable time.

Special thanks to Dr.Dhiman Dutta, Head of the Department of Statistics, Asutosh College. I am greatly indebted to Dr. Sankha Bhattacharya, Dr. Parthasarathi Bera and Dr. Shirsendu Mukherjee. Faculty members often took pain and stood by me in adverse circumstances. Without their encouragement and inspiration it was not possible for me to complete this project.

Finally my earnest thanks go to my friends who were always beside me when I needed them without any excuses and made these three years worthwhile. This project is not only a mere project. It is the memories spend with the whole department which has created a mutual understanding among us. There are many emotions related to this piece of work, especially respect and duty towards teachers and vice versa, educational attachment with my friends and social attachment with my college.

· Declaration:-

I, Sohini Bhadra, a student of B.Sc. Sem-VI, Statistics Honours, of University of Calcutta, Registration No. 012-1211-0539-19; Roll No. 193012-11-0287 hereby declare that I have done this piece of project work entitled as "Food-crops Production Of India and agricultural growth – A Case Study and Data Analysis" under the supervision of Oindrila Bose Assistant Professor, Department of Statistics, Asutosh College, as a part of B.Sc. Sem-VI examination according to the syllabus paper DSE-B2.

I further declare that the piece of project has not been published elsewhere for any degree or diploma or taken from any published project.

Appendix

Necessary calculations:

>adf_test <- read_excel("adf test.xlsx")

> View(adf_test)

> attach(adf_test)

> install.packages("tseries")

> library(tseries)

>plot.ts(GDP)

```
> adf.test(GDP)

        Augmented Dickey-Fuller Test


data:  GDP
Dickey-Fuller = -0.48987, Lag order = 3, p-value =
0.9788
alternative hypothesis: stationary
```

here p value is greater than 5%level of significance .hence null hypothesis cannot be rejected. Hence our data is non –stationary. A unit root is present in the data.

```
>library(tseries)
> dgdp=diff(GDP)
> adf.test(GDP)

Augmented Dickey-Fuller Test
data:  GDP
Dickey-Fuller = -0.48987, Lag order = 3, p-value =
0.9788
alternative hypothesis: stationary

> adf.test(dgdp)

        Augmented Dickey-Fuller Test
data:  dgdp
Dickey-Fuller = -2.9107, Lag order = 3, p-value =
0.2097
alternative hypothesis: stationary
```

```
adf.test(RICE)

        Augmented Dickey-Fuller Test
data:  RICE
Dickey-Fuller = -2.9083, Lag order = 3, p-value =
0.2099
alternative hypothesis: stationary

> y=diff(RICE)
> adf.test(y)

      Augmented Dickey-Fuller Test
data:  y
Dickey-Fuller = -5.8603, Lag order = 3, p-value =
0.01
alternative hypothesis: stationary


> adf.test(WHEAT)
Augmented Dickey-Fuller Test
   data:  WHEAT
Dickey-Fuller = -2.7313, Lag order = 3, p-value =
0.281
alternative hypothesis: stationary

> z=diff(WHEAT)
> adf.test(z)

        Augmented Dickey-Fuller Test
data:  z
Dickey-Fuller = -4.6925, Lag order = 3, p-value =
0.01
alternative hypothesis: stationary

> adf.test(CEREAL)

        Augmented Dickey-Fuller Test
data:  CEREAL
Dickey-Fuller = -0.70051, Lag order = 3, p-value =
0.9644
alternative hypothesis: stationary

> x=diff(CEREAL)
> adf.test(x)

        Augmented Dickey-Fuller Test
data:  x
Dickey-Fuller = -4.8748, Lag order = 3, p-value =0.01
```

```
alternative hypothesis: stationary

> adf.test(PULSES)

        Augmented Dickey-Fuller Test
data:  PULSES
Dickey-Fuller = 0.14631, Lag order = 3, p-value =
0.99
alternative hypothesis: stationary

> w=diff(PULSES)
> adf.test(w)
Augmented Dickey-Fuller Test
data:  w
Dickey-Fuller = -5.2197, Lag order = 3, p-value =
0.01
alternative hypothesis: stationary
>lmtest::dwtest(GDP)
> model1=lm(GDP~RICE+WHEAT+CEREAL+PULSES)
> summary(model1)

Call:
lm(formula = GDP ~ RICE + WHEAT + CEREAL + PULSES)

Residuals:
      Min         1Q     Median         3Q        Max
-5.217e+11 -1.704e+11 -1.386e+10  1.309e+11  6.388e+11

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.964e+12  2.738e+11  -7.171 5.68e-09 ***
RICE        -2.216e+09  1.016e+09  -2.182  0.03437 *
WHEAT        3.029e+09  9.770e+08   3.100  0.00333 **
CEREAL       3.975e+09  1.484e+09   2.678  0.01030 *
PULSES       8.660e+09  2.495e+09   3.471  0.00116 **

Residual standard error: 2.688e+11 on 45 degrees of freedom
Multiple R-squared:  0.9066,  Adjusted R-squared:  0.8983
F-statistic: 109.2 on 4 and 45 DF,  p-value: < 2.2e-16
> lmtest::dwtest(model1)

        Durbin-Watson test

data:  model1
DW = 1.2371, p-value = 0.0007564
alternative hypothesis: true autocorrelation is greater than 0
```