Analysis of Indian Census Data

- Project Report : Group 27 -
- Data Mining : CS685A -

Anupriy	150121
Deepanshu Bansal	150219
Kshitij Jaiswal	150340
Pranjal Gupta	150506
Vipul Yadav	150811

Indian Institute of Technology Kanpur
Department of Computer Science and Engineering

Introduction

Data Analysis plays an integral role in the development of any country and public availability of huge census data of India encouraged us to analyze the various factors contributing to the success and some others hindering the progress of nation in various aspects. The Indian Census is the most credible source of information on Demography (Population characteristics), Economic Activity, Literacy and Education, Housing Household Amenities, Urbanization, Fertility and Mortality, Scheduled Castes and Scheduled Tribes, Language, Religion, Migration, Disability and many other socio-cultural and demographic data since 1872. This is the only source of primary data in the village ,town and ward level.

First of all, why is census important ?

- Equitably distributing the billions of dollars of public money requires up-to-date population data. Census helps with the fair and impartial distribution of public funds among various departments such as federal and state funding for educational programs, health care, law enforcement, highways and new infrastructures etc is allocated based on population.
- It provides valuable information for planning and formulation policies for Central and the State Governments and is widely used by National and International Agencies, Scholars, business people, industrialists, and many more.
- The Delimitation/reservation of Constituencies- Parliamentary/Assembly/Panchayats and other Local Bodies is also done on the basis of the demographic data thrown up by the Census. Census is the basis for reviewing the country's progress in the past decade, monitoring the on going Schemes of the Government and most importantly, plan for the future. That is why the Slogan is **"Our Census - Our Future"**.

So, we have analyzed and studied the data available in following fields to prove or disprove some hypotheses and to give some useful insight and suggestions that should be taken by the government for the betterment and continuous growth of the country :

- First we have done analysis of road accidents in India and various parameters associated with it viz. effects of weather conditions, defective vehicles etc. Also the relation between funds allocated by government in this and number of accidents is also analyzed. We present extensive analysis of this problem and its possible solutions.
- Second we have done extensive analysis of crimes done and reported. These include crimes done against children/overall, sex offences, expenditure on central police to reduce these crimes etc. This is done as this is the one which contribute the most in the hindrance of development and growth.
- Third we have estimated growth of country in upcoming years by analyzing the agricultural data of past years.
- Last we also tried to prove Kaldor's hypothesis of growth.

Road Accidents Analysis

Analysis of factors affecting road accidents is of utmost importance to reduce frequency of occurrence. We highlight hypothesis of data of road accidents of various parts of India. Non-parametric Chi Square method is applied to carry hypothesis of data and to conclude the factors which promote accidents. Our hypothesis includes the study of relationship between various types of locations and year wise accidents, occurrence of accidents and persons killed in different states of India, fatalities occurrence frequency in different years and study of relationship between time, weather and place of accidents.

1. ***Null hypothesis is made on the basis that year wise accidents and type of weather are independent. Alternative hypothesis is assumed that year wise road accidents and type of weather are not independent.***

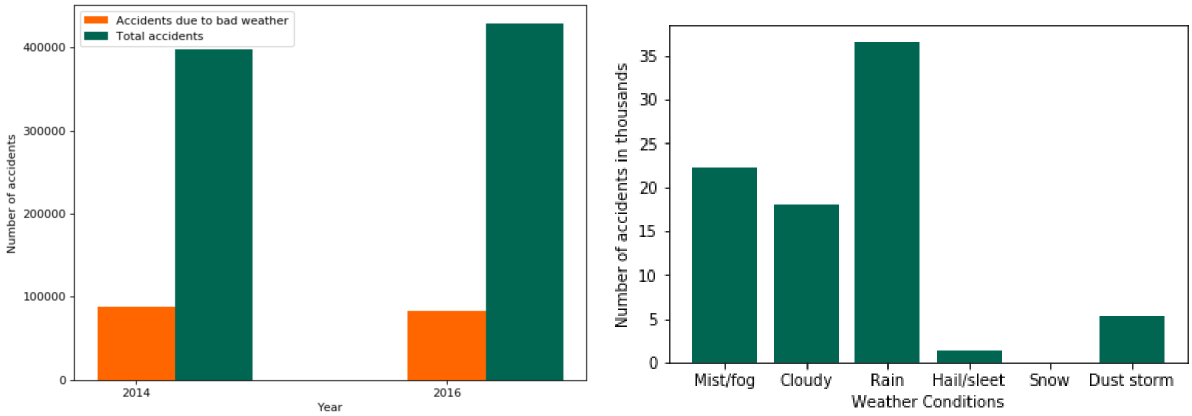

Figure 1: Left figure shows the comparison of total accidents and accidents due to bad weather. Right figure shows distribution of accidents due to different weather conditions.

   Null hypothesis is rejected as our chi square value is 4588. which is higher than the standard chi-square values at 1%, 5% and 10% level of significance for degree of freedom 4. It is observed that approx. 25% of the total accidents are because of bad weather.

   1 out of 4 road accidents are because of bad weather

2. ***Null hypothesis is made on the basis that year wise accidents and vehicular defects are independent. Alternative hypothesis is assumed that year wise road accidents and accidents due to vehicular defects are not independent.***

   Null hypothesis is rejected as our chi square value is 5043 which is higher than the standard chi square values at 1%, 5% and 10% level of significance. So, year wise road accidents and type of vehicular defects are not independent as we can also see from the figure 2 that as the number of accidents increases, the proportion of number of accidents due to defective break also increases.

   We found that among all the factors which cause accidents due to vehicular defects, defective breaks is most common characteristic among these accidents which is also evident from the pie chart shown in figure 3.

   Number of accidents due to defective brakes increased from 51.3% in 2014 to 63.9% in 2016 despite of decrease in total accidents by 12.9%. We predict that if it continues then this rate will increase to 76.5% in 2018.

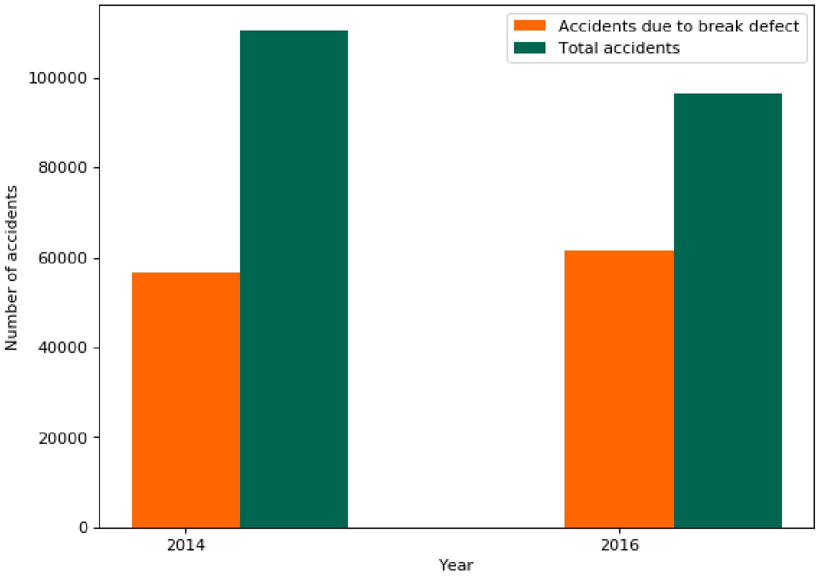

Figure 2: Total accidents and accidents due to Break defect

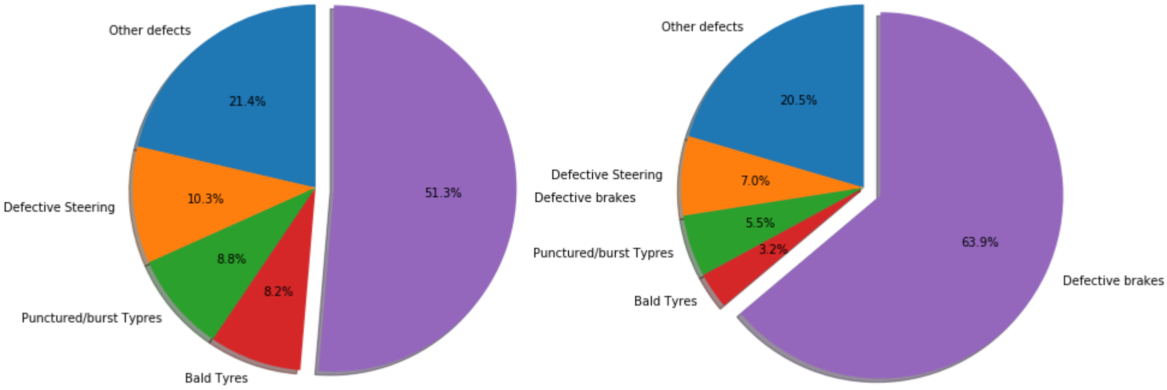

Figure 3: Distribution of accidents due to Vehicular Defects. Left fig. is for 2014, Right fig. is for 2016

3. ***Null hypothesis is made on the basis that total number of accidents in a state due to condition of roads and funds allocated for roads to that state are independent. Alternative hypothesis is assumed that number of accidents in a state due to condition of roads and funds allocated for roads to that state are dependent.***

   Chi square value for contingency table 1 comes out to be 31.1 and from chi square table it is lower than value for 1% significance level. So, By chi square test it is observed that number of accident on a particular road condition is dependent on fund allocated to that road conditions at 5% and 10% significance level but independent for 1% significance level.

4. **Suggestions after above observations :** Broadly dividing the causes of accidents in three categories of accidents due to road conditions, due to vehicular defects and due to weather conditions we observe that most accidents are happening because of bad road conditions which are around 71%.

Table 1: Contingency Table for frequency of states

No of Accidents(K)/Budget	0-58 Cr.	58 - 116 Cr.	116-174 Cr.	174-232 Cr.	232-290 Cr.
0-67	20	3	0	1	1
67-134	0	2	3	0	0
134-201	1	0	1	0	0
201-269	1	0	1	0	0
269-336	0	0	1	1	0

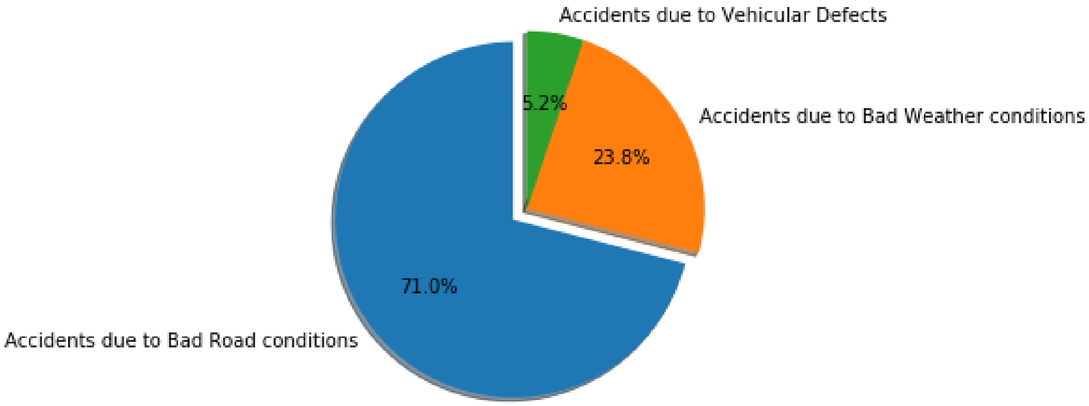

Figure 4: Distribution of total accidents due to various reasons

So among all the funds allocated we propose that government should spend more budget in improving the conditions of roads and should frame policies to regulate the safety guidelines strictly in the automobile industry to lower down the accidents happening because of vehicular defects.

Crime Analysis

To a developing nation the ingredients to constant and fast growth are a must to have. Thus the search to find relation between different factors that contribute to its growth has to be done. In this region we will look into how Crime, measures taken by Government to curb the crimes and the GDP of the nation. The dependency of a prosperous nation on the peace prevailing there and its GDP is presumed here to be directly related.

1. **Sex Ratio during education vs Sex Offences**
   GPI is a measure of relative access to education of males and females. This index is released by UNESCO. In its simplest form, it is calculated as the quotient of the number of females by the number of males enrolled in a given stage of education. We have taken this as a measure of how does education to both the sexes affects the sexual offences committed by them in the future.

   - **Hypothesis :** ***When the GPI approaches 1 we expect the number of crimes conducted as sexual offences should decrease.***
   - **Assumptions :** Sexual offences committed in the age of 18-30 years in year x are done by the children of age 11-13 years in the year $(x - 10)$
   - We have analyzed the GPI data of 1991-1999 to the sexual offences in the years 2004-2012. We have four categories of sexual offences i.e. Rape, Dowry Deaths, Molestation and Cruelty by husband or relative of husband.

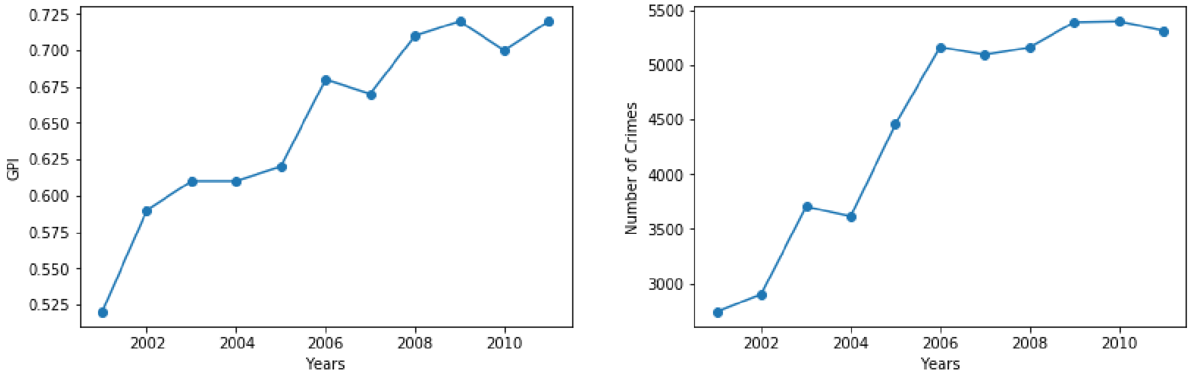

Figure 5: First figure shows the comparison GPI vs Years. Second figure shows number of sexual offences vs Years

   - **Observations :**
     - Over the course of time GPI tends to reach its optimal condition i.e. 1
     - The crime rates are not increasing at a higher rate from 2006 onwards as they were before.
     - The number of sexual offences stagnates as the GPI reaches 1.
   - We build the regression model to analyze the sexual offence crime rate variation with the GPI. From the figure 6 it might seem that as the GPI increases to 1 number of crimes are also increasing but it is not the case as the population is also increasing. To compare them more accurately we build the table 2.
   - We have used the regression model to predict the population in year 2024. From the data we have GPI of 1 in year 2011 and its effect according to our hypothesis

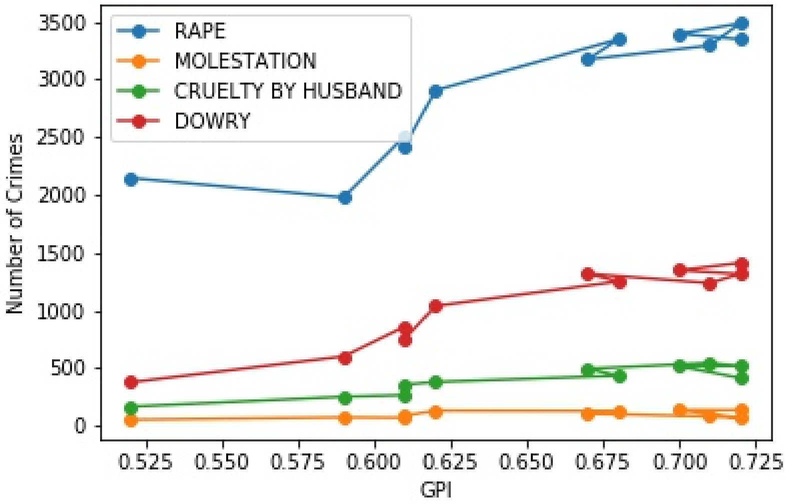

Figure 6: Sexual Offences vs GPI

Table 2: Year vs Total Population

Year	Total Population(in crores)
2011	121
2024	146

will be visible in year 2024. We compare the GPI of 0.72 in year 1998 and its effect in year 2011 with GPI of 1 in year 2011 with its effect in year 2024.

Table 3: Per Person offence (PPO) predicted for year 2024

Sexual Offence	#crimes in 2011	PPO total ($x10^{-7}$)	#crime in 2024	PPO ($x10^{-7}$)
Rape	3349	27.67	4651	31.85
Molestation	136	1.12	46	0.31
Cruelty By Husband	415	3.42	378	2.58
Dowry	1413	10.85	1476	10.10

- We observe that almost all the sexual offences have decreased with GPI increasing to 1. So we propose that government should spread awareness to educate children of both sexes so that the GPI will remain 1 and thus leading to lesser number of criminal sexual offences.

2. **Percentage change in Crime vs GDP**

- **Hypothesis :** *An organization/institution must prosper when the crime rate there is low*
- **Assumptions :**
  - The growth of a State is synonymous to its GDP.
  - The cases filed in that year is taken synonymous to the crimes committed in that duration.

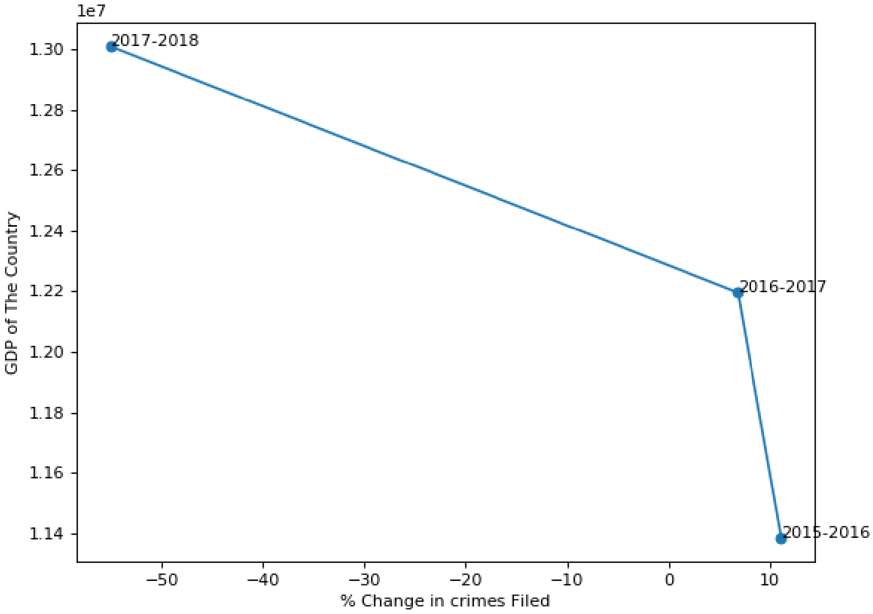

Figure 7: GDP vs the % change in the crime cases filed

- **Result :**
  - Figure 7 pertains to the relationship between percentage change in crimes filed from the period between 2015 to 2018 and the respective change in GDP over this period. Here we can clearly see that whenever crimes reported go down a rise in the corresponding years GDP is observed. Thus the hypothesis that nations prospers whenever crimes are low is thus proved.

3. **GDP vs Expenditure on Central Police**

- **Hypothesis :** ***An organization/institution must prosper when they invest in controlling the crimes/atrocities committed in their jurisdiction***
- **Assumptions :**
  - The growth of a State is synonymous to its GDP.
  - The Expenditure on Central Police is presumed to cause effect on the GDP of the succeeding year. This presumption is taken since it will take some time to invest the allocated fund for proper use.
- **Result :**
  - Here we observe from the graph in figure 8 that government is continuously spending on Central Police Forces like AR, BSF, CISF, CRPF, ITBP, NSG, RPF, SSB, etc. and growth in GDP over time. The observation of these two being directly proportional proves the hypothesis that the expense on Central Police improves the GDP of the nation in time.

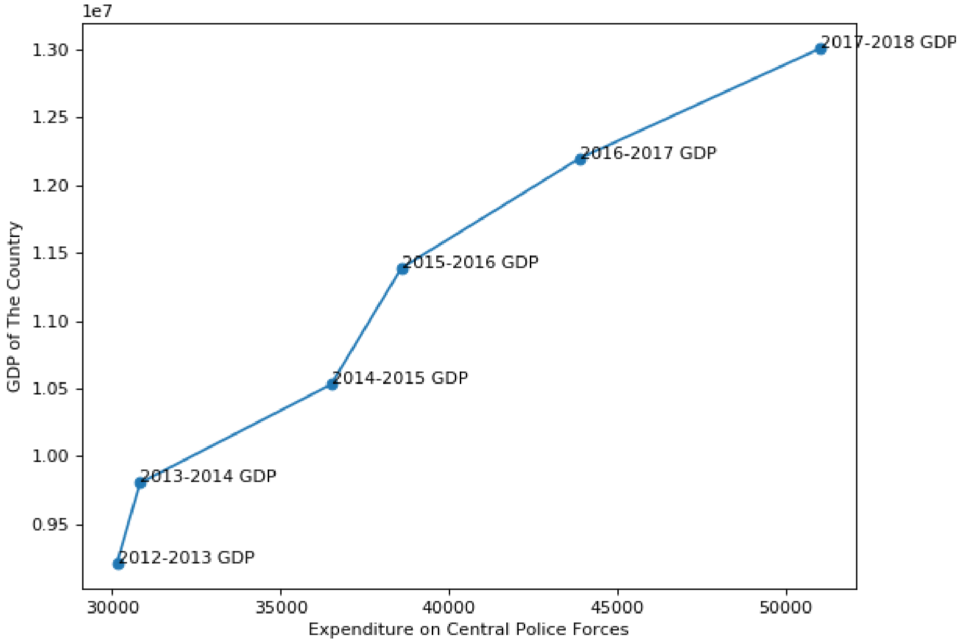

Figure 8: GDP vs Expenditure on Central Police

4. **Crimes vs Expenditure on Central Police Forces**

- **Hypothesis :** ***As the expenditure on Central police forces increase crimes filed in country should decrease***
- **Assumptions :**
  - Expenditure on Central Police Forces in an year is reflected after some duration as implementation and awareness takes time. We've made an assumption of 2 years lag.
  - For the year 2018, data for crime filed is only for initial 6 months. So over the year, we've just doubled the numbers.

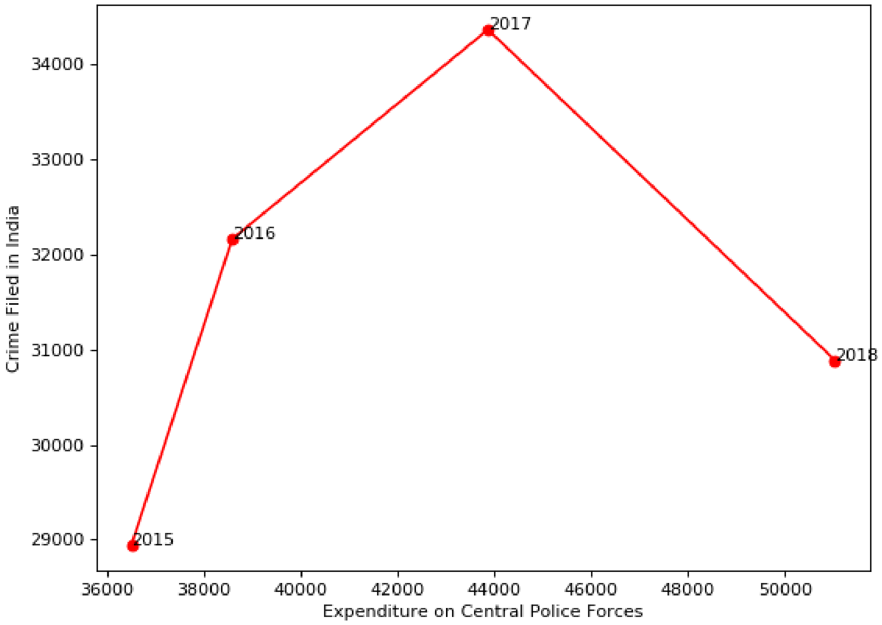

Figure 9: Crimes vs Expenditure on Central Police

- **Observations :**
  - As expenditure increases from year 2013 to 2015, crimes filed is also increased from year 2015 to 2017 which is opposite to our hypothesis. But in the following year of 2018, number of crimes filed has increased.
  - In a way, we can say that decrease in number of crimes filed is due to the cumulative expenditure on central police forces

Impact of Inputs on Agricultural Productivity

India is the world's seventh largest country and has the world's second largest population. It ranks second worldwide in farm outputs. Agriculture and allied sectors accounted for 13.7% of the GDP in 2013 and provides employment to 50% of the countries workforce.. As the country's economy is growing, the contribution of agriculture in its GDP is also steadily declining. But, the fact is that agriculture is the broadest economic sector and plays a significant role in the overall socio-economic fabric of India. On the other hand, Indian agribusiness is as yet confronting the issues, for example, low level of business sector reconciliation and integration, availability of dependable and convenient information needed by farmers on different issues in farming.

Aim:

Estimate the growth in productivity for the coming years.

Assumptions:

We have assumed that relation between growth in various inputs to get the growth in agricultural productivity is linear.

Work Done:

We know that the agricultural productivity growth is a linear function of growth of inputs. So,

$$productivity\_growth = f(land\_growth, labour\_growth, capital\_stock\_growth)$$

We have found this using a linear regressor.

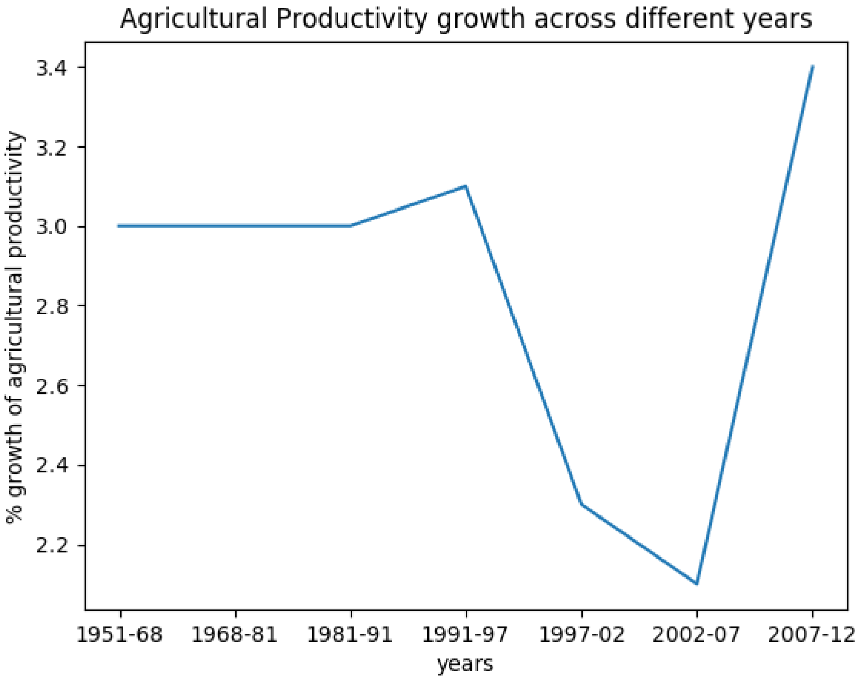

Figure 10: Agricultural productivity growth across different years

Now, to get the values of inputs for the upcoming years, we fitted a polynomial curve in the input-year curves and extrapolated them to get the corresponding values of inputs.

From the graph of land growth and years, we found that the relation between them is

$$productivity\_growth = -0.008(year\_index)^3 + 0.131(year\_index)^2 - 0.627(year\_index) + 1.219$$

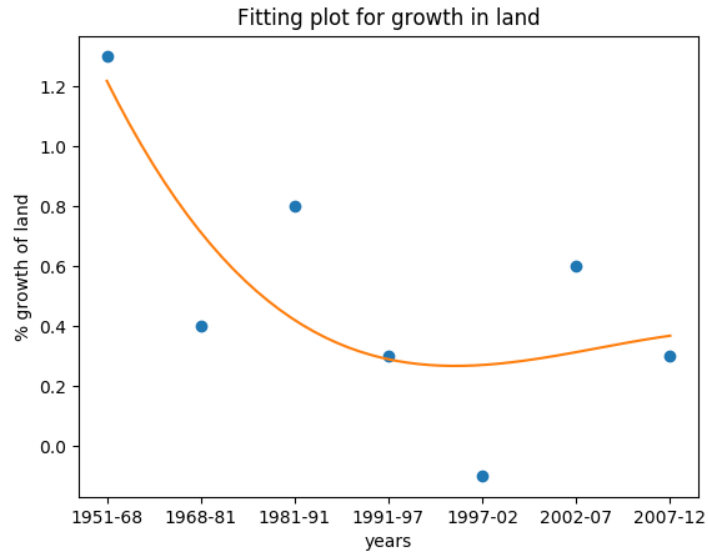

Figure 11: Land growth across different years

From the graph of labour growth and years, we found that the relation between them is

$$productivity\_growth = 0.058(year\_index)^2 - 0.4321(year\_index) + 1.681$$

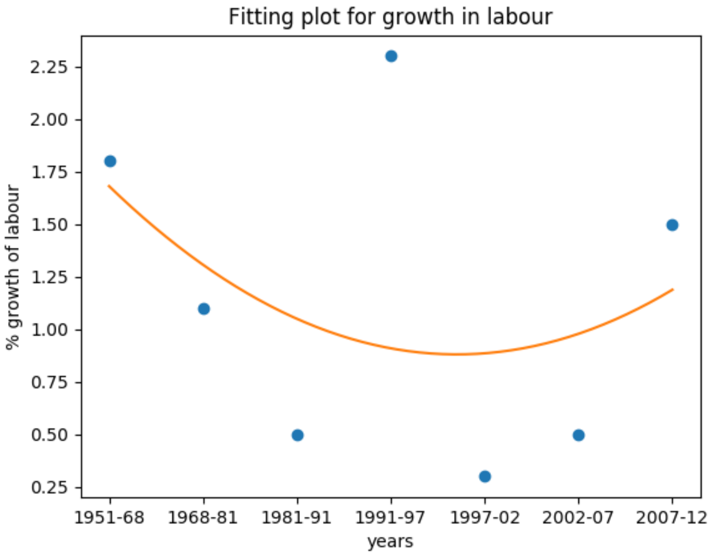

Figure 12: Labour growth across different years

From the graph of fixed capital stock growth and years, we found that the relation between them is

$$productivity\_growth = 0.125(year\_index)^2 - 0.2536(year\_index) + 2.836$$

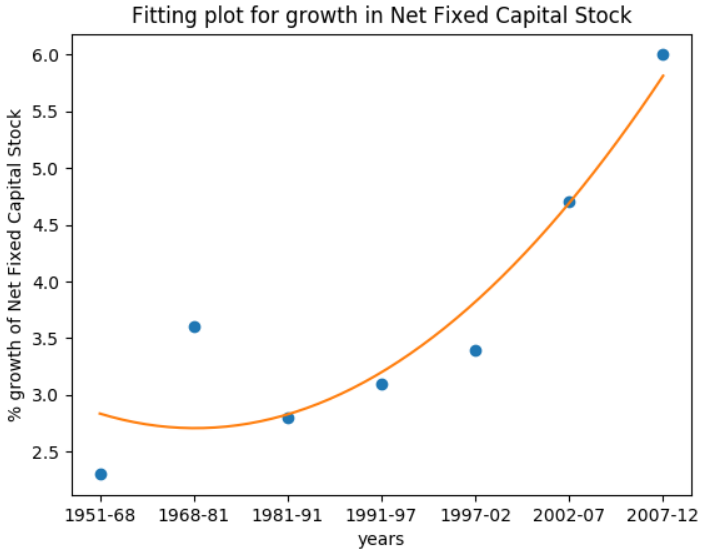

Figure 13: Fixed capital stock growth across different years

Results:

- The land growth, labour growth and capital stock growth for the next plan is found to be 0.385, 1.514, 7.185 respectively.
- The production growth for the next year is estimated to be 3.193 .

Kaldor's Hypothesis of Growth

Kaldor's hypothesis states that faster the rate of growth of manufacturing output, the faster will be the rate of growth of GDP. It is also known as manufacturing-as-the-engine of economic growth hypothesis.

We have used State Development Product (SDP) as a measure of growth of GDP for various states and then plotted the growth of manufacturing output and growth of SDP for the financial years 2008-09, 2009-10 and 2010-11. We observe for some states manufacturing growth rate is negative while corresponding SDP growth rates are positive.

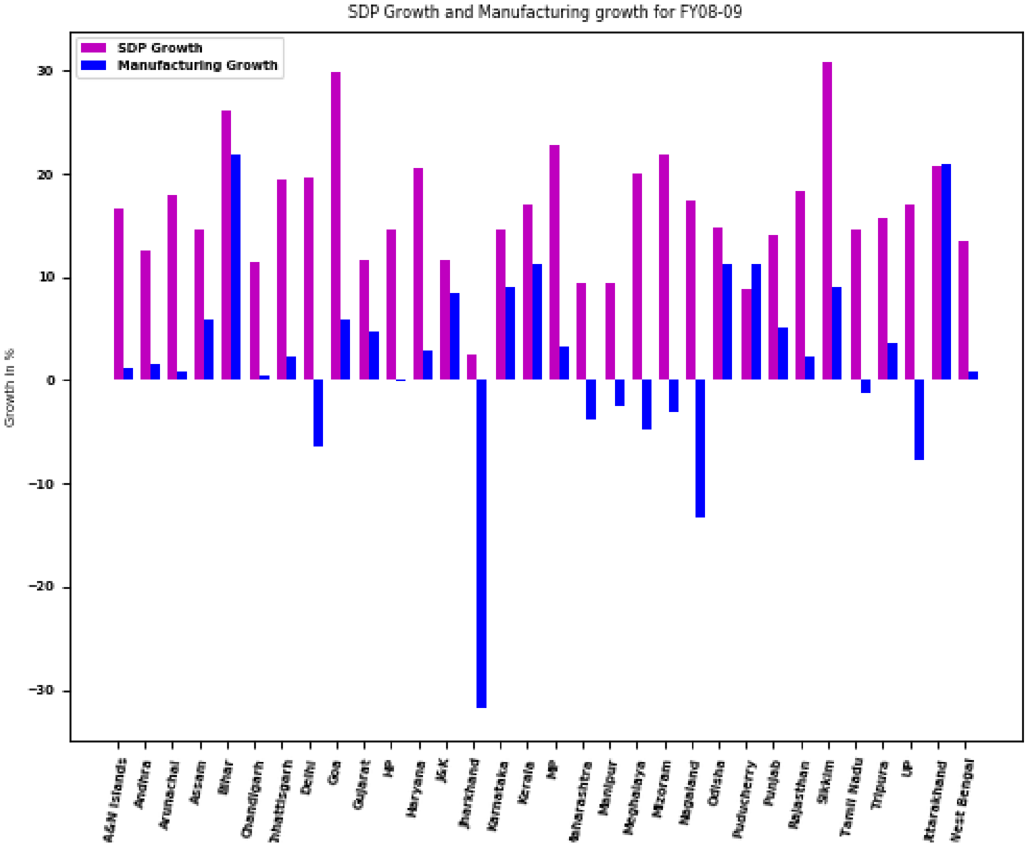

Figure 14: State Development Product (SDP) growth and Manufacturing Growth for FY08-09

We observe the following graphs of Average growth of SDP vs Average Manufacturing growth over the years for the states Nagaland, Uttar Pradesh (and many more) which have inconsistent growth of SDP in relation to the growth in manufacturing output. For these states Kaldor's hypothesis is not valid as increasing manufacturing output, growth rate of GDP is not increasing. But Average growth of SDP vs Average Manufacturing growth over the years for the states Jharkhand, Andhra Pradesh we can say that Kaldor's hypothesis is valid. Kaldor's hypothesis is rejected for the smaller period of time in indian states.

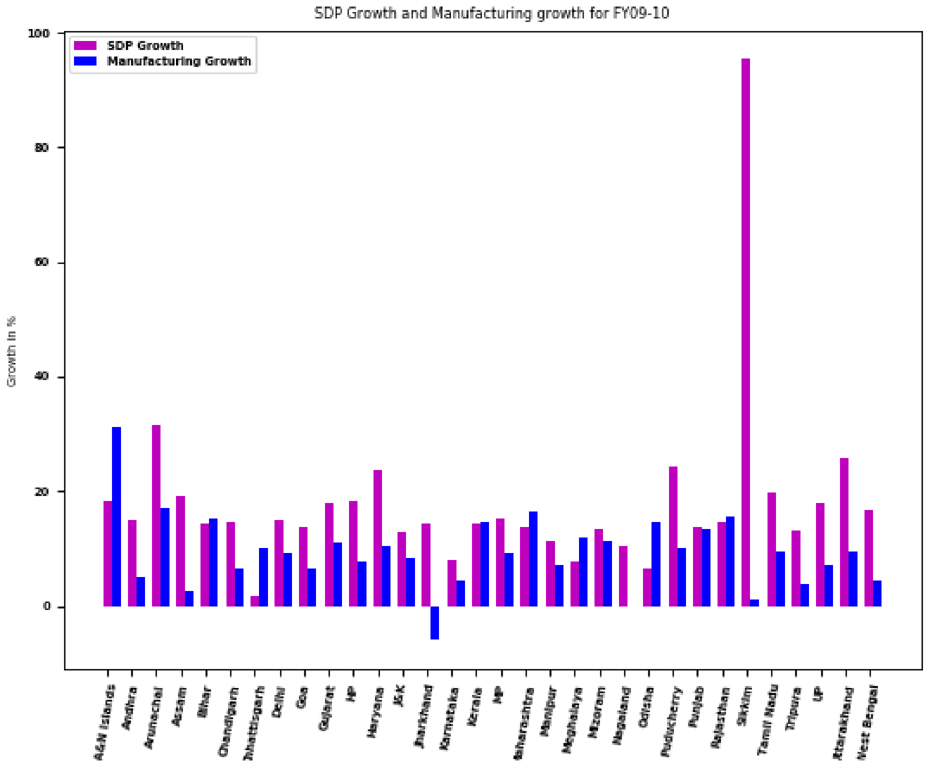

Figure 15: State Development Product (SDP) growth and Manufacturing Growth for FY09-10

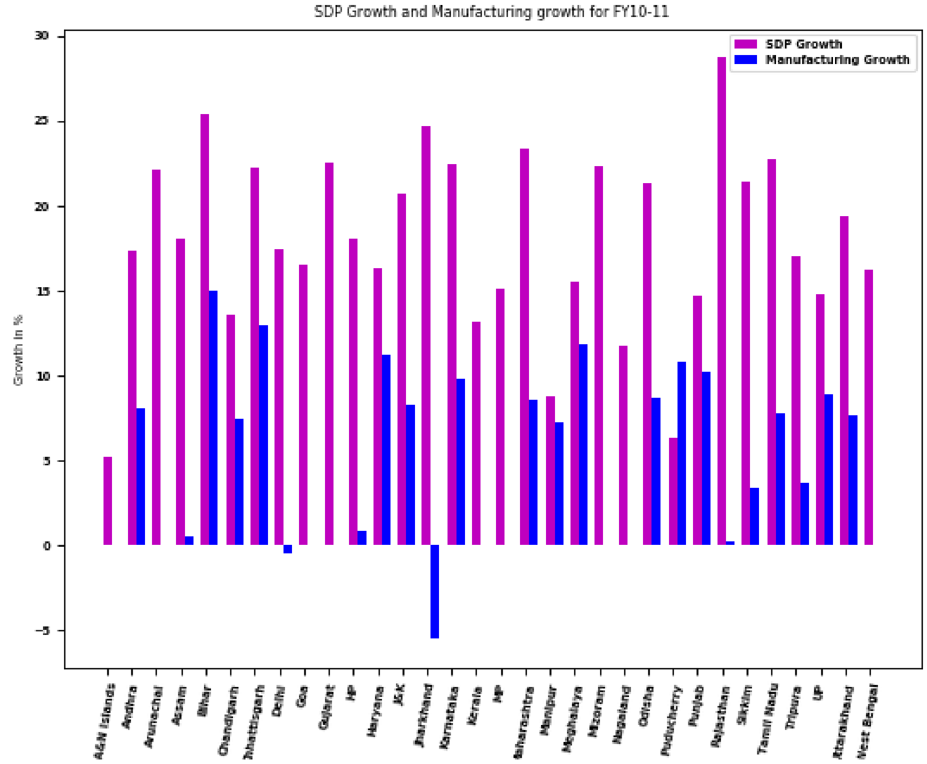

Figure 16: State Development Product (SDP) growth and Manufacturing Growth for FY10-11

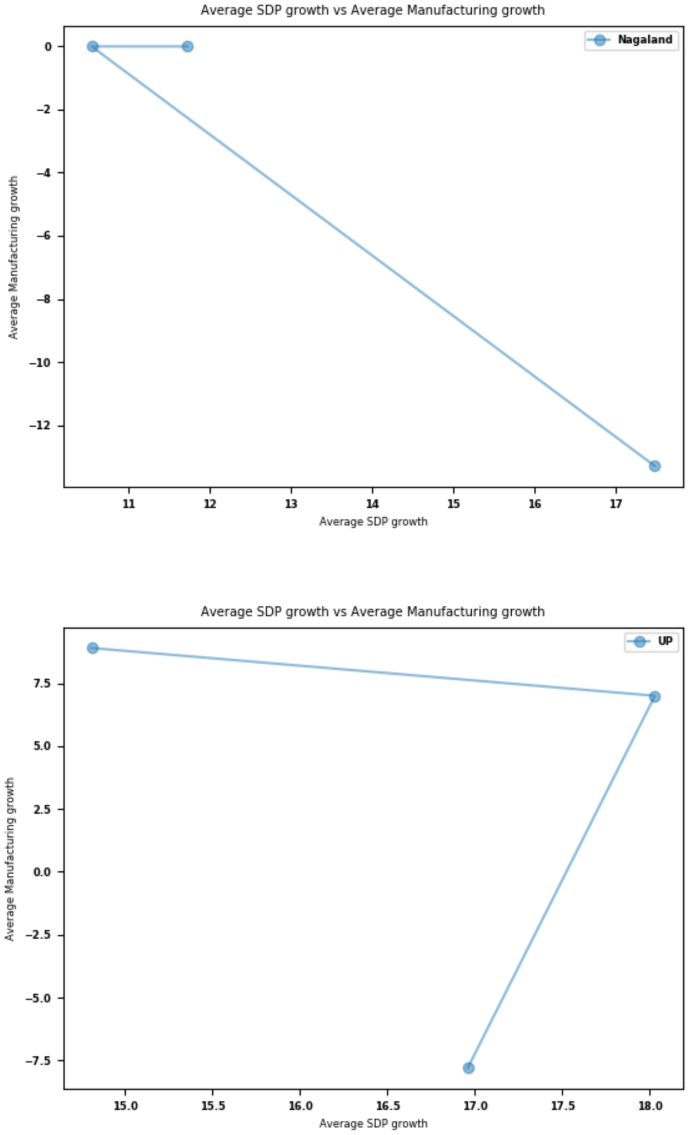

Figure 17: Average SDP growth vs Average Manufacturing growth for the states with negative growth rate

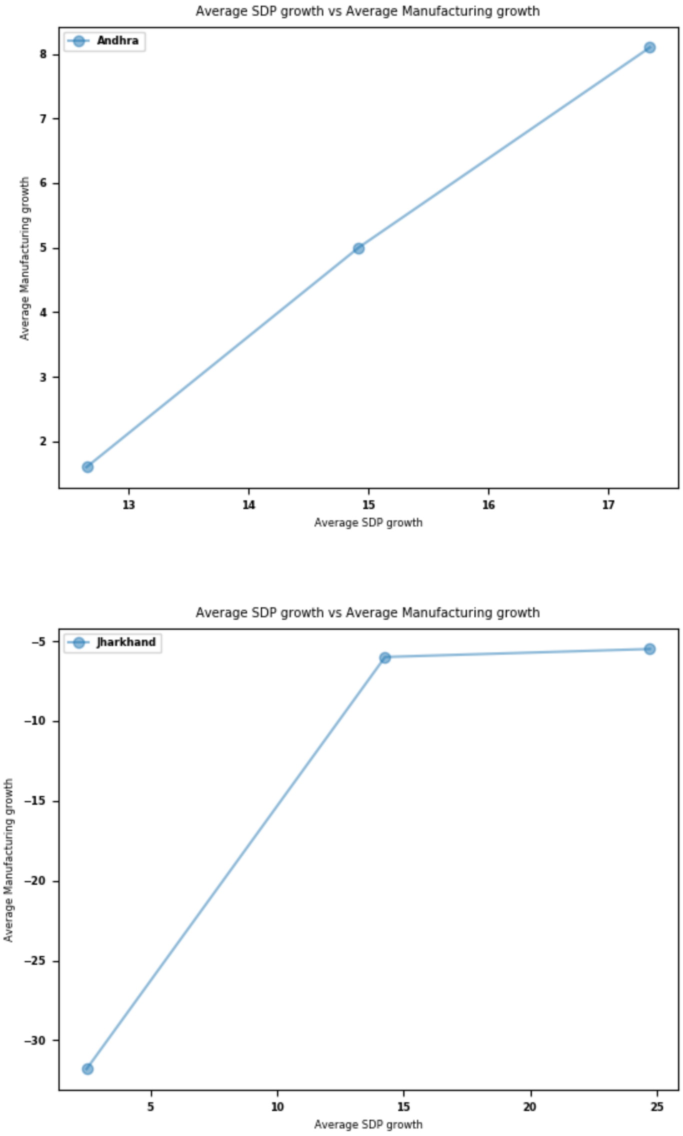

Figure 18: Average SDP growth vs Average Manufacturing growth for the states with positive growth rate

Dataset

- Open Government Data Platform India.
- Indiastats

References

- *Pooja Sikdar et al.* Hypothesis of data of road accidents in India-review.
- *Rahmi Yamak et al.* A Re-Examination of Kaldor's Engine-of-Economic Growth Hypothesis for the Turkish Economy
- *Panchanan Das* Economic Reform, Output and Employment Growth in Manufacturing: Testing Kaldor's Hypotheses
- *John Adams et al.* Determinants of Agricultural Productivity in Rajasthan, India: The Impact of Inputs, Technology, and Context on Land Productivity

GitHub Link

Code - `https://github.com/anupriy97/Analysis-of-Indian-Census`