RAW CROP REPORT

Finding correlation between literacy, rainfall, groundwater quality with crop production across India

Biswarup Das Sarma
Ganesh Halthota
Girish Naik
Harsha Pamarthi Vardhan

11.12.2021
DA - 231 Mtech (Online)

Img src: https://www.euractiv.com/section/agriculture-food/special_report/sr-agri-24-feb-how-is-tech-revolutionising-the-agricultural-sector/

Repo: https://github.com/ganesh737/iisc-da231-raw_crop

INTRODUCTION

Correlation between literacy, rainfall, groundwater quality with crop production across India. It is intuitive that good rainfall is conducive to good production. How do the regions perform during poor rainfall in the rainy season? We also want to cross match with the literacy rate of the region to see if higher levels of literacy affects crop production.

HYPOTHESIS

We expect the timely and adequate rainfall to positively affect crop production. We expect production to increase with higher literacy in rural areas.

DATASET

Dataset	Source	Size	Format	Granularity
Crop/Agriculture	Kaggle: Agriculture Produce India	2,46,091 X 7 (~14MB)	CSV file	States and Districts in India. Years - 2000-2014
Water Quality	CPCB Site	~1500 records x 18 columns * 8 years	Individual HTML for each year	Detailed upto individual water sources for Years 2012-2019
Rainfall	Kaggle: Rainfall in India	4116 records X 19 columns	CSV file	States in India. Years - 1901-2015
Literacy	Govt. of India Literacy Rate	36 records x 8 columns	CSV file	State level from 2001 and 2011 census

APPROACH

We will be using PySpark DFs and MatplotLib on Google colab for computation and analysis. Data-sets to be used from various official sites - kaggle, pollution control board mentioned above. To find the correlation between literacy, rainfall, groundwater quality with crop production, we find the answers to the below mentioned questions.

1. What is the trend of top 5 crops per region in each season for the years 2000-2014
2. How has rainfall affected the production of kharif crops for the years 2000-2014?
3. Effect of Literacy on crop production for each state in the years 2001, 2011-12
4. For the crops Arecanut, Jowar, Jute, Rice show the performance of these crops across states for the time periods 2000-2014 when the rainfall is above or below average.
5. How does water quality affect the crops of a region across the years 2012-2014?

Data Readiness

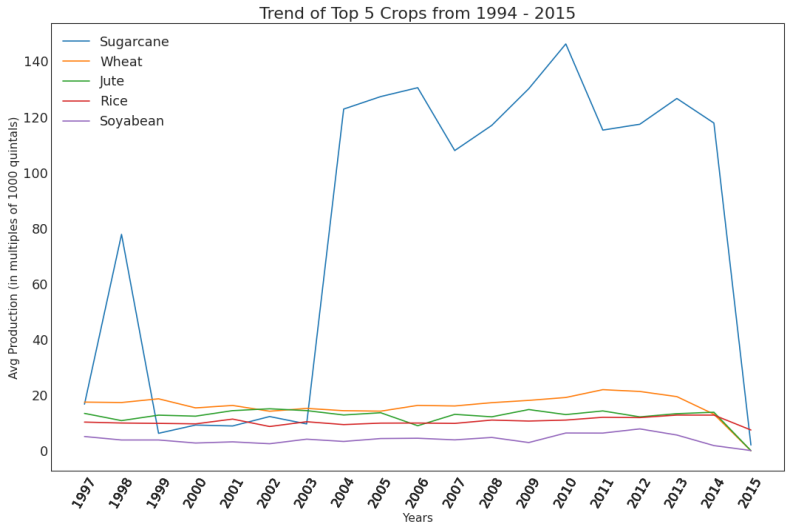
Load the multiple csv files for Crop Dataset and Preparation of all data into usable format(csv) for ground water data set.
The crops produced vary by region and local weather conditions(water, rainfall, etc). So the analysis will be performed by grouping the states into zones of North, South, East and West.

Individual High Level Approaches

1. Top 5 Crops.
    a. Perform filter to remove crops which have less data for the years 2000-2014.
    b. Sum the produce for each region
    c. Get the top 5 crops per region.
    d. Show a appropriate graph
2. Correlation between Crop production and for Kharif crop.
    a. Filter on the years 2000-14 for Rainfall and Kharif crops.
    b. Sum the crop production and rainfall recorded over regions.
    c. Join them on year and region.
    d. Correlation between one year rainfall and crop production to its previous year.
3. Correlation between Literacy with crop produce.
    a. Filter the crops data by the years 2001 and 2011
    b. Normalise the crops data and literacy data so that crop production and literacy values lie between 0-100
    c. Find the divergence between literacy and crop production for both the years
    d. Find the average divergence and average literacy
    e. Plot the above two data points and analyse for any correlation between them,
4. Correlation of the crops Arecanut, Jowar, Jute, Rice with Rainfall in Rainy season
    a. Filter on the years 2000-14 for all datasets.
    b. Filter the crop dataset for the crops Arecanut, Jowar, Jute, Rice
    c. Find the average rainfall for all the states.
    d. Find the produce of individual crops vs rainfall and plot the graph.
    e. Find the correlation between rainfall and crop production.
5. Water Quality Index vs Agricultural Efficiency
    a. Perform filter to remove crops which have less data for the years 20012-2014.
    b. Load the multiple csv files for water quality for the years 2012-2014.
    c. Calculate the Agricultural Efficiency.
    d. Calculate the Water Quality index.
    e. Merge the two tables.
    f. Find the correlation.

OBSERVATION

Top 5 Crops



During this analysis we find that below are the top 5 crops which are high in production:

1. Sugarcane
2. Wheat
3. Jute
4. Rice
5. Soyabean

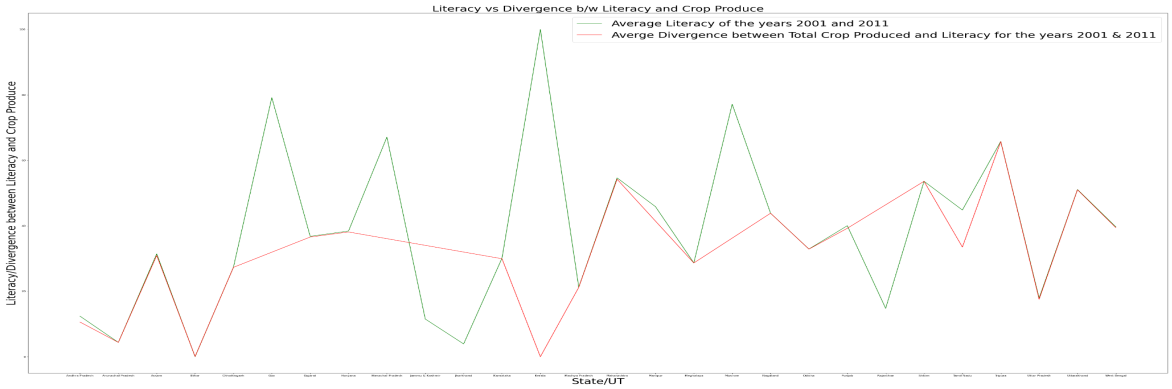
Literacy vs Crop Produce

We observed data for two years – 2001 and 2011 for which literacy data was available through census dataset. We want to observe the effect literacy of a region has on the production of crops. Does higher literacy, which can mean an increased level of industrialisation in farming, translate to better crop produce.

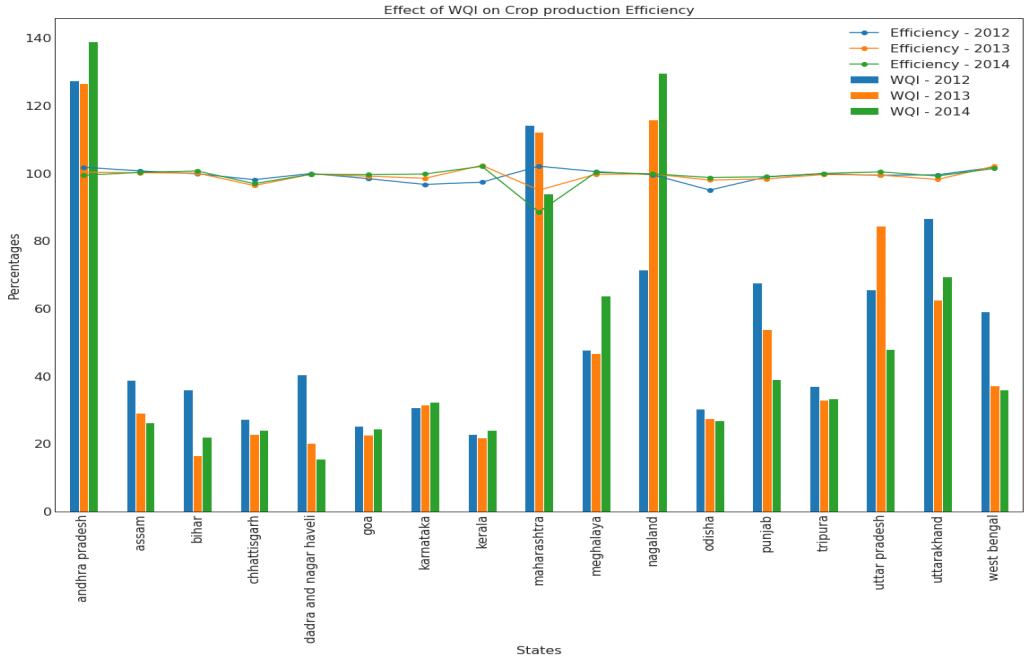
The plot has two lines:

1. Line 1 – average literacy for the years 2001 and 2011. It is normalised to lie between 0-100
2. Line 2 – average of divergence of literacy and crop produce for the two years. Both crop produce and literacy are normalised between 0-100. Divergence is calculated as the absolute difference between the two values

We plotted the two lines together to look for correlation between them.



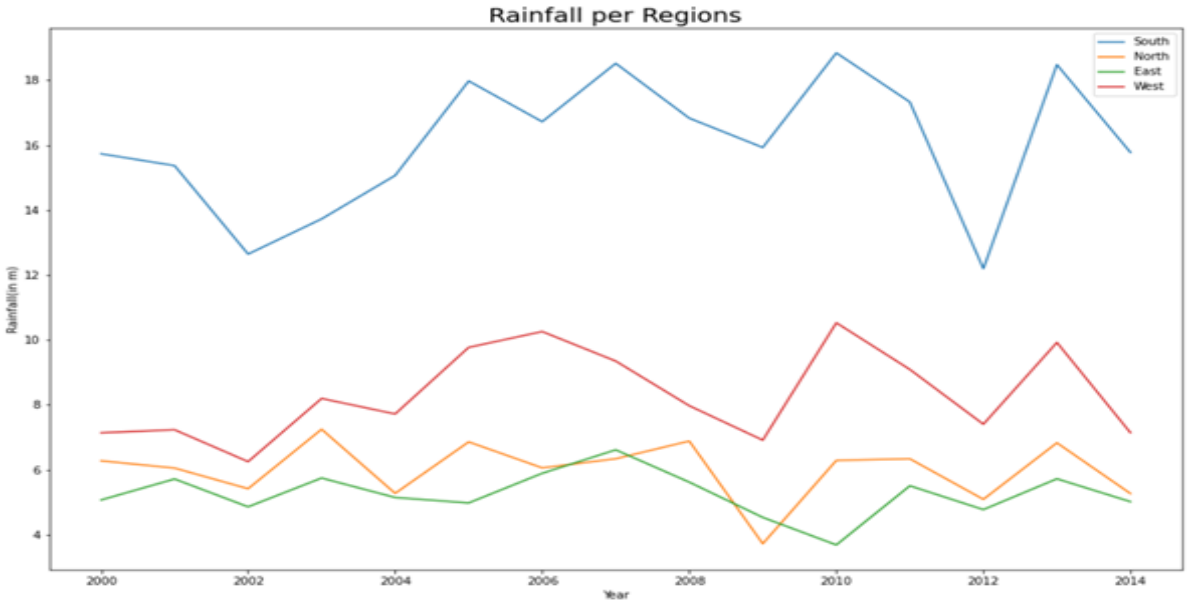
Water Quality Index vs Agricultural Efficiency



During this evaluation we have found that water quality has very little effect on the crop production efficiency.

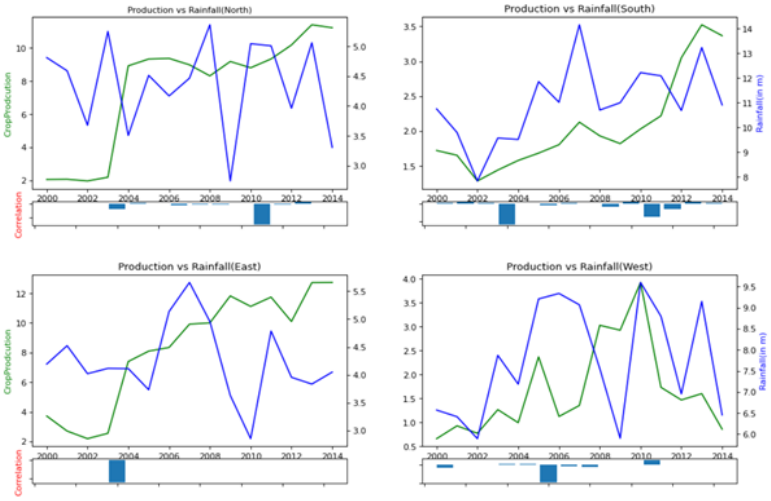
Rainfall vs Crop Produce

The plot has four lines for 4 zones.



The plot has four subplots, each for one region. Each subplot contains two parts, one for crop production, rainfall, and another for correlation between them. Below are lines

with bars in it.



1. **Line 1** – Total Crop Production per Total Area in each region

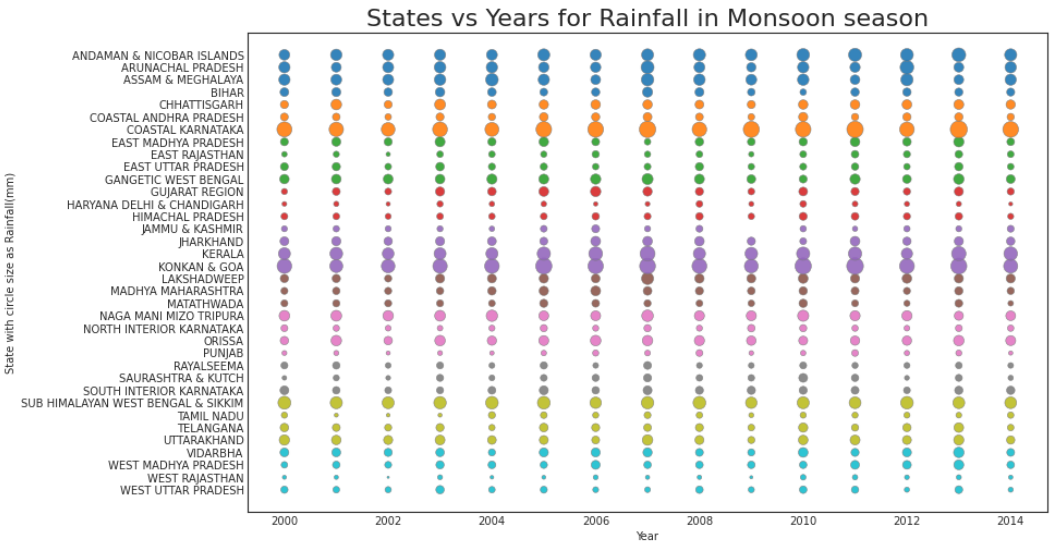
2. **Line 2** – Total rainfall in that region in meters.

3. **Bar** – Correlation value between crop production and rainfall.

Comparison of Arecanut, Jowar, Jute, Rice Performance with Rainfall in Rainy Season

The concentration of production of the crops Arecanut, Jowar, Jute, Rice were evaluated to understand the typical weather conditions of the region.

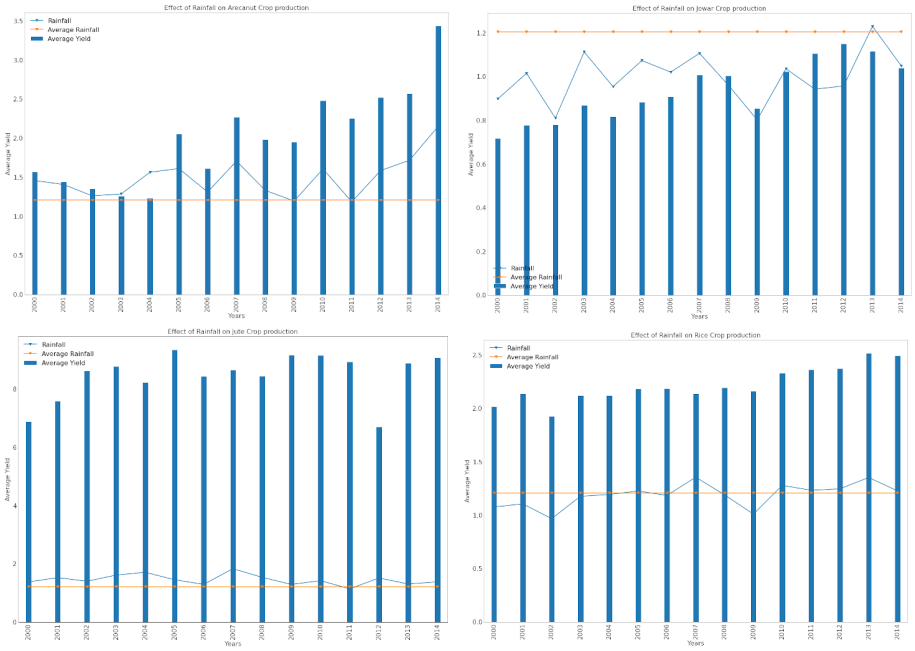
The rainfall varied as below over the years for the states in India -

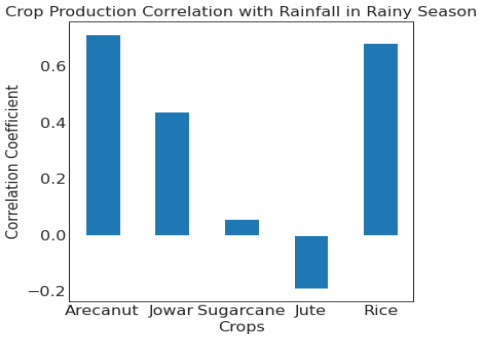


We see that some of the states have higher rainfall compared to others and can support water intensive agriculture compared to other states.

The below plot shows that some of the crops are concentrated in certain areas.

Based on the rainfall during Monsoon season in the respective regions, their production varies as below -



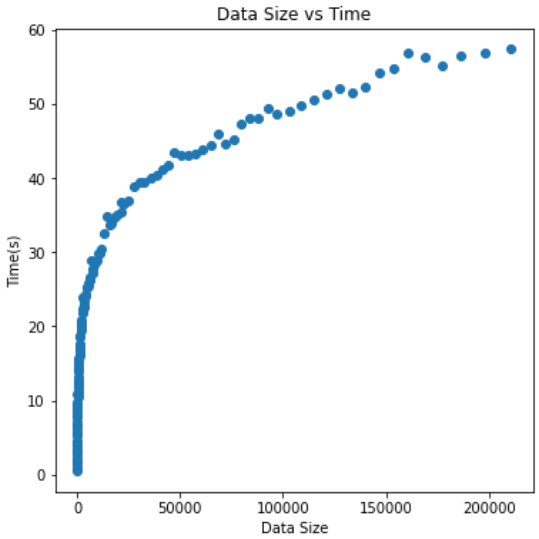


The correlation between these crops and rainfall was checked. The plot is as shown on the left. For some crop we can observe that

- Arecanut and Rice show strong positive correlations to rainfall in the rainy season.
- Sugarcane shows moderate positive correlation to rainfall in the rainy season

- Jute shows inverse relation to rainfall in the rainy season

EVALUATION



Evaluation is performed on the Google Colab platform. The Python notebook is evaluated for timing based on the time taken for transformation with actions, including conversion from spark dataframe to pandas for reporting the final results. It is found that the overall evaluation takes *~6 mins 15 sec*.

In order to see the scalability of the code, we have considered the code for crop performance with rainfall. It was executed for 124 cycles with varying data sizes and the distribution of time taken against data size is as right.

From the plot, the time taken follows a logarithmic curve of data size. So the complexity of the curve approximates to O(log(n)).

Additional time for evaluation with each new crop for filtering is dependent on the data set size available for the crop. We found that the addition of each new crop resulted in additional evaluation time of *~0.35s* in the best case (~10 additional records) and ~3s in the worst case(~12000 additional records).

CONCLUSION

Based on our study, we conclude that

When literacy shoots up, the divergence dips sharply or at-least does not rise. High divergence indicates higher unrelatedness between the crop produce and literacy. Hence higher literacy can contribute to better crop produce.

We have also found that water quality has very little effect on the crop production efficiency.

When rainfall increases, crop production also increases in most cases. There are few cases where there is sudden dip and might be because of rainfall at the end of Kharif season. When compared to all regions, the North zone has less correlation to rainfall.

For some crop we can observe that

- Arecanut and Rice show strong positive correlations to rainfall in the rainy season.
- Sugarcane shows moderate positive correlation to rainfall in the rainy season
- Jute shows inverse relation to rainfall in the rainy season

CHALLENGES AND GAPS

1. Water data set from 2015 is in pdf format and unstructured, hence not used for comparision.
2. It would have been more realistic to analyse this data had the literacy data been available for all the years instead of 10 years apart.
3. Some of the state's crop production data was abnormally high for these two years, maybe due to previous years correction or some unknown reasons

REFERENCES

1. For WQI:
    a. https://link.springer.com/article/10.1007/s13201-017-0579-4
    b. http://www.indiaenvironmentportal.org.in/files/water%20quality%20index.pdf
    c. https://www.youtube.com/watch?v=HTkNmmMoUzE
    d. https://www.youtube.com/watch?v=LtXfIYYb8F4
    e. http://iitk.ac.in/iwd/wq/drinkingwater.htm
2. For Agricultural Efficiency:
    a. https://www.youtube.com/watch?v=MB_K_a0xo3g
3. Understanding Correlation Coefficient - https://www.scribbr.com/statistics/correlation-coefficient/