Effect of groundwater quality and rainfall and agricultural output

Musthyala Rushabh
B.E Computer Science and Engineering
BITS Pilani, Hyderabad Campus

Sistla Shashank
B.E Chemical Engineering (Hons) with Minor in Data Science
BITS Pilani, Hyderabad Campus

Varma Vivek
B.E Computer Science and Engineering
BITS Pilani, Hyderabad Campus

Abstract— ***India is primarily an agrarian country. This paper presents an analysis of the relationship between crop yield, ground-water quality and rainfall using various techniques, with special emphasis on the year 2014 (due to availability of data). Ground-water quality judged on pH, Temperature, Biological Oxygen Demand, Fecal Coliform and Total Coliform.***

Keywords—data analysis, clustering, correlations, rainfall, agricultural output

I. OBJECTIVES

1) Data Preprocessing –
   a) Removing NaN values in all 3 datasets (rainfall, agricultural output, groundwater quality) using appropriate methods respectively.
   b) Binning data according to time periods to help reduce the numerosity and improve interpretability.
2) Water Quality analysis –
   a) Clustering the states based on their groundwater quality to understand how geographical location could play a role in determining certain characteristics of the water
3) Simliar analysis on rainfall and agricultral output
4) ependence of agricultural output on rainfall and groundwater quality
   a) Conduct a season by season analysis on which states produce the most crops and look for possible reasons to explain the above based on the rainfall and groundwater quality in the region under consideration.
   b) Provide graphs to support the above claims

II. BACKGROUND

Ground water quality affects the agricultural output. The data is mined in order to draw out which qualities of the groundwater most significantly influence the crop yield across seasons.
First, the data is used in order to test the hypothesis that water quality has an effect on crop yield, and if the hypothesis is true, the correlations developed can help broaden the understanding of the methods which can help reform current agricultural methods, leading to increased crop yields.

III. INITIAL DATA ANALYSIS

In this section, we walk through each dataset, explaining how we tackled the issue of NaN values and then our analysis of the dataset in question.

A. Groundwater Dataset

This dataset has 24 columns, 21 of which have numeric data. The 21 themselves are the mean, minimum and maximum values of seven features, which is spread across 18 states. First, the dataset is consolidated by concatenating all the datasets of each state. In this dataset, there were a total of 1836 missing values. Stochastic Regression Imputation was chosen to deal with the missing values.

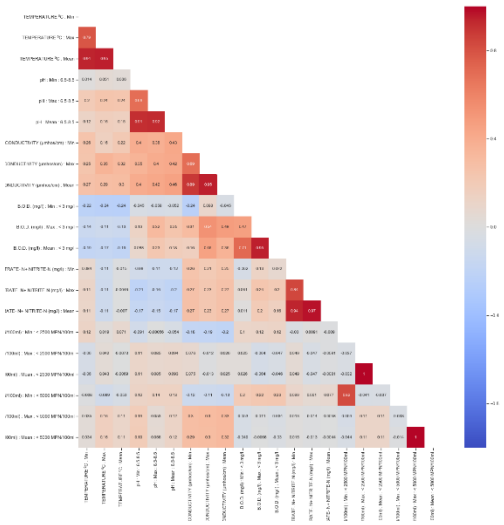
Stochastic Regression Imputation: [1]

Stochastic regression imputation is implemented via three steps:

1) Mean Imputation
2) Regression Imputation
3) Stochastic Regression Imputation

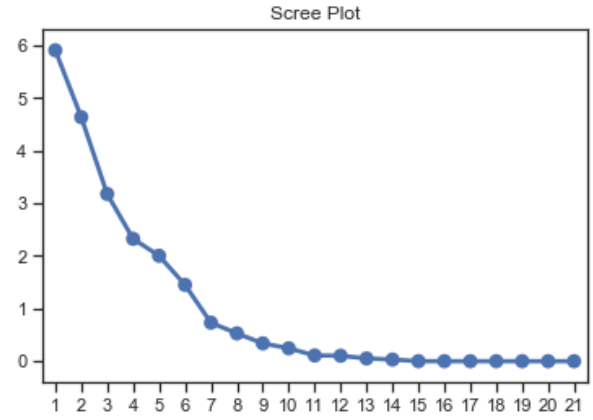
Mean imputation fills in the missing values with the mean of the corresponding feature. This is required since there is more than one missing value per record. We can't impute more than one feature at a time using regression, thus, to regress for a missing feature, columns of placeholder features are added via mean imputation. Second, Regression Imputation is carried out to fill in the missing values one by one, using the placeholder mean imputed values whenever necessary. However, regression imputation has a drawback. The newly filled points lie on the regression hyperplane, this implies that the correlation between the predictor variables and the predicted variables is unity. To account for this, Stochastic regression imputation takes care of the bias by augmenting the predicted variable by a random variable, which follows a normal distribution with mean zero and variance equal to the variance of the residual data.

After the dataset was cured of missing values, PCA was used to reduce the dimensionality of the data. PCA should only be used when there is correlation between the features. [1]
As each feature had three columns explaining the nature of the feature, there is correlation between the features, as shown in this heat map visualization of correlation.



Heat Map

To select the principal components, components with eigenvalue greater than 1 were picked. A total of 7 principal components were obtained, as clearly interpretable from the graph above, where 7 triplets of high correlation are observed. This number also lies in accordance with the scree plot obtained:

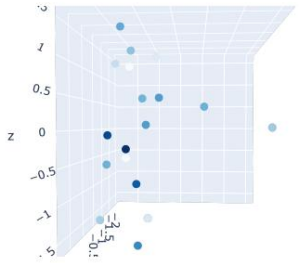


Scree Plot

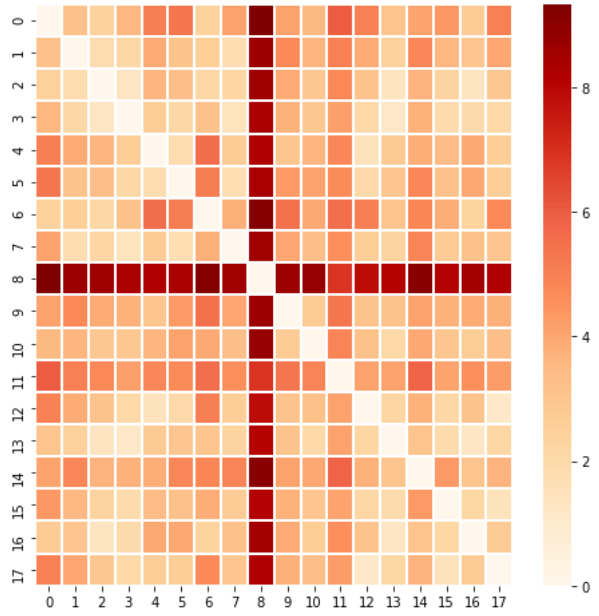
In this case, we can more or less identify what the significance of each of the principal components is. In order to make use of the PCA data, we took the average of all the records state-wise in order to cluster the states.

The dataset has columns pertaining to different metrics it comes to when determining groundwater quality (temperature, pH, Biological Oxygen Demand, Fecal Coliform and Total Coliform).

We started the clustering process by plotting a 4D graph of the reduced data and computing the distance matrix to gain some insight and intuition as to how our clusters would turn out.



4D graph of PCA groundwater data

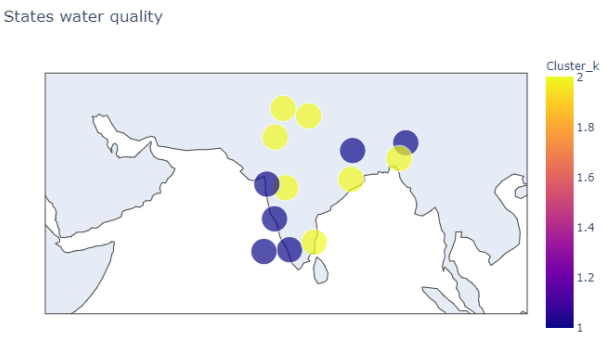


Distance Matrix

Both K-means clustering [2] and DBSCAN [3] were implemented on the obtained data and we found the clusters found by DBSCAN to be in line with the information given by the 4D plot and distance matrix.

The clusters obtained were as follows –

Noise: Bihar, Chhattisgarh, Lakshadweep,
Cluster: Andhra Pradesh, Assam, Daman Diu Dadra Nagar Haveli, Goa, Himachal Pradesh, Kerala, Madhya Pradesh, Kerala, Maharashtra, Odisha, Pondicherry, Punjab, Rajasthan, Tripura, Uttar Pradesh, Uttarakhand, West Bengal
For better interpretability, we plotted this data on the India map to have a visual representation of how groundwater varies across the country.



Geographical visualization of clusters

After gaining some understanding as to how groundwater varies across the country, we felt the need to understand the interrelationships between the different properties. To do this, first binned the data (equal frequency) into 3 bins - low, medium and high and ran the Apriori algorithm [4] on it to generate association rules with a confidence of above 95%.

The association rules were as follows –

1) High Conductivity, High Total Coliform levels --> Medium pH

2) High Total Coliform levels, Medium Nitrate/ite conc. --> High Fecal Coliform levels

3) High Temperature, High Total Coliform levels --> Medium pH

4) High B.O.D, High Fecal Coliform levels, Medium pH --> High Total Coliform levels

5) High Conductivity, High Fecal Coliform levels, High Total Coliform levels --> Medium pH

6) High Fecal Coliform levels, Low Nitrate/ite conc., Medium pH --> High Total Coliform levels

With the following characterization of medium (values below the range are low, above are high)

Feature	**Medium Range**
B.O.D	(1.0, 2.71]
CONDUCTIVITY	(611.333, 1298.0]
FECAL COLIFORM (MPN/100ml)	(35.667, 1012.451]
NITRATE-N+NITRITE-N	(0.5, 2.0]
TEMPERATURE ºC: Mean: <5000 MPN/100ml	(25.0, 28.0]
TOTAL COLIFORM	(20.0, 3434.741]
pH Mean: 6.5 – 8.5	(7.2, 7.6]

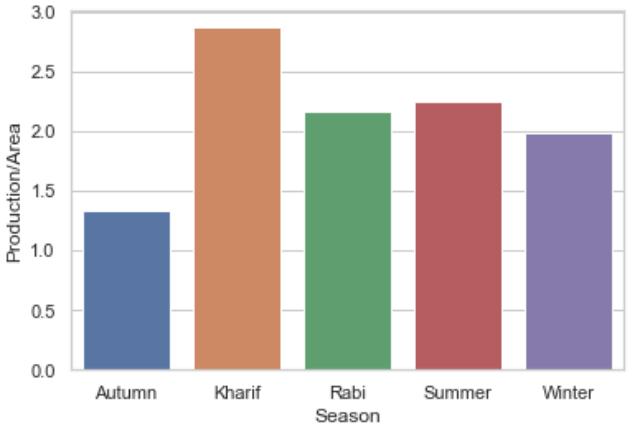
B. Agriculture Dataset

This dataset has large numerosity with unnecessary segregation of the data. In this dataset, binning was implemented and the missing values were filled with average produce per unit area for that particular crop. This data was made more useful and computable by combining the yields of the crop, both district-wise and season-wise, in bins of size 3 years.

It's important to note that India has the following agricultural seasons - Kharif, Rabi, Winter, Summer and Autumn. There is some overlap between these seasons but the crops have been classified as such. The time period of the seasons is as follows –

1. Kharif: June - September
2. Rabi: October - February
3. Winter: December - February
4. Summer: April - June
5. Autumn: September - November

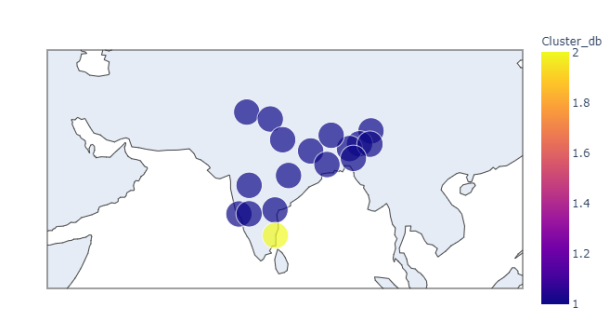
The agricultural output season wise was plotted to get a better understanding of the data.



Agricultural output per season

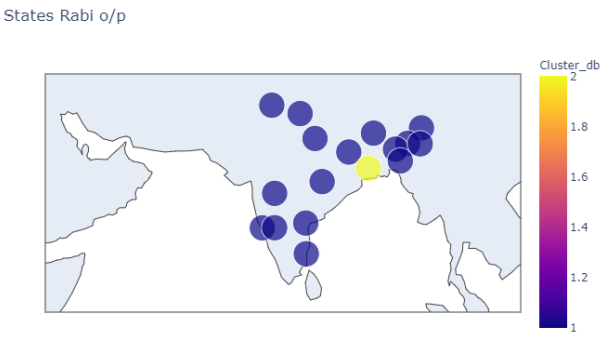
Given the availability of groundwater and rainfall data for the year 2014, we decided to cluster the states based on their agricultural output in that year, season wise. Throughout this process, the clusters given by both K means and DBSCAN were identical (unless specified). Since each state had a different amount of land dedicated for agriculture which put a current glass ceiling on their output, we decided to cluster based on Production/Area to make for a fair comparison.

- In the Kharif season we observed Puducherry to have a significantly higher value of agricultural output



Geographical visualization of Kharif ssn

❖ In the Rabi season, it was found that West Bengal had a significantly higher value than the remaining states



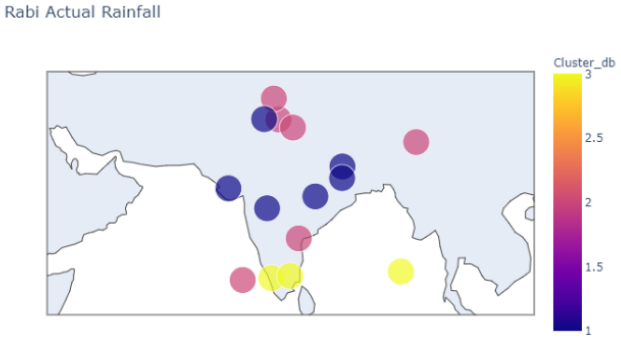
Geographical visualization of Rabi ssn

C. Rainfall Dataset

Similar to the approach used for the previous datasets, we performed K-means clustering on the rainfall data on a season by season basis to understand how different states experience rainfall differently throughout the year.

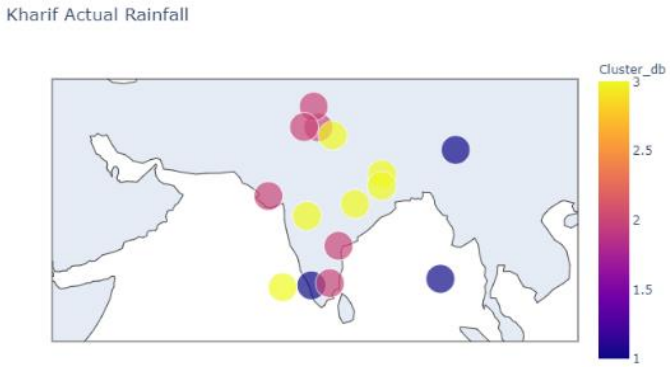
We saw more distinct trends in the Kharif and Rabi seasons, which is understandable given how those are the major agricultural seasons in India.

❖ In the Rabi season we observed a distinct difference in the actual rainfall levels based on how north/south a particular state is. Most north states fell into one cluster, south states in another and the central states in the 3rd cluster.



Geo. visualization of Rabi ssn rainfall 1

❖ In the Kharif season we observed more of a west vs east separation as compared to the north-south comparison earlier



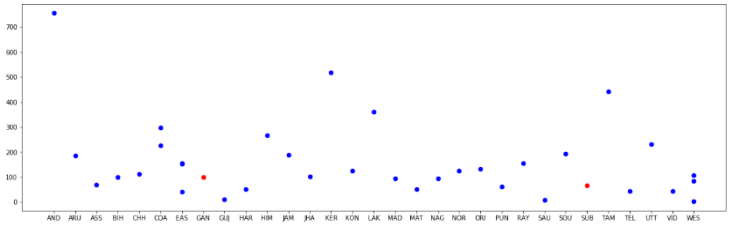
Geo. visualization of Kharif ssn rainfall

IV. RELATIONSHIPS BETWEEN RAINFALL AND GROUNDWATER VS AGRICULTURAL OUTPUT

With the data now preprocessed analyzed and clustered, we are in a position to now gather relations and examine how these affect each other.

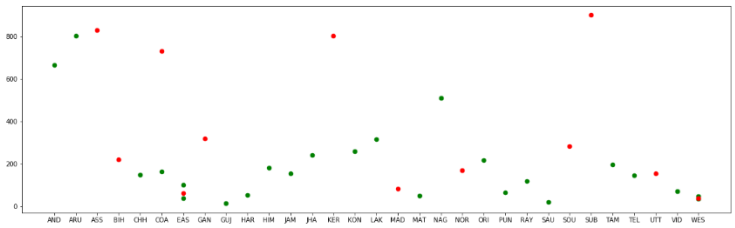
A. Rainfall vs Agricultutural Output

With regards to rainfall, we observed that West Bengal lies on the lower end of the spectrum with regards to actual rainfall and it has the highest value of production per area. The data gives us inferences consistent with the knowledge that Rabi crops to well in states with less rainfall. (The outlier state(s) are colored in red).



Rainfall during Rabi season (WB in red)

In summer, we observed that many of the states that recorded large numbers in terms of production per area also recorded a lot of rainfall during the summer months, in accordance to the knowledge that summer crops tend to require a larger amount of rainfall. (The outlier state(s) are colored in red).



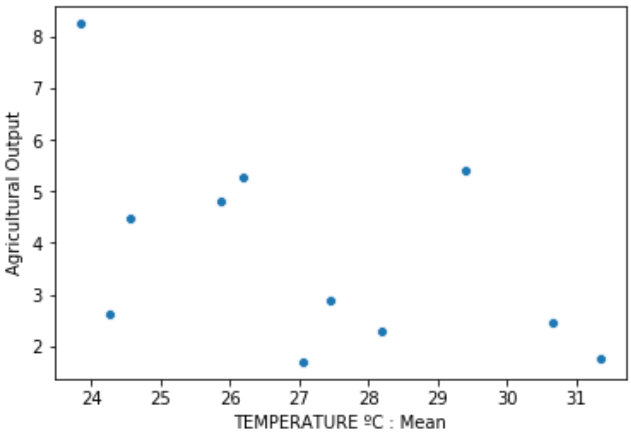
Rainfall during summer season

B. Groundwater vs Agricultural Output

For the properties of groundwater, we took them individually and plotted them against the agricultural output and

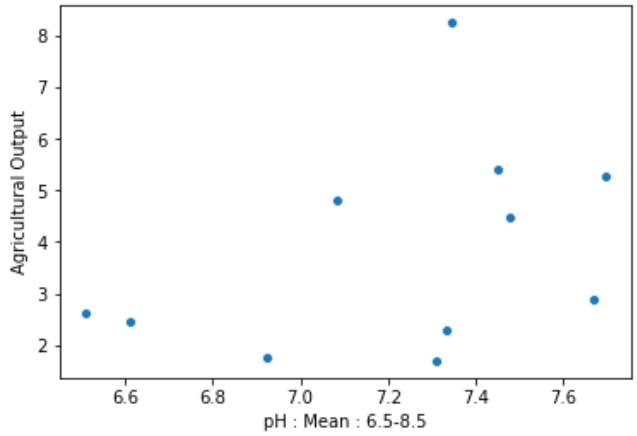
calculated the Pearson coefficient to help us understand the correlation between the features.

The 2 most interesting results came from the temperature and pH of the water. The temperature vs agricultural output graph had a Pearson coefficient of -0.58 indicating a moderate negative correlation between the two.



Temp vs Agri. O/P

We also saw that the graph between pH and agricultural output gave us a Pearson coefficient of 0.43. From these, we can say with a bit of confidence that lower water temperatures and higher pH are advisable for growing crops.



pH vs Agri. O/P

V. CONCLUSION

Throughout our study, we implemented a variety of techniques ranging from stochastic regression imputation and clustering to the Apriori algorithm to help us understand the intricacies of groundwater and rainfall and furthermore, the role they play on the agricultural output all over our country.

Our studies have helped us understand the relationship between things like B.O.D and conductivity in groundwater, how they can influence each other. Furthermore, we have proved with more certainty the reliance of Rabi crops on a large amount of rainfall and how Kharif crops tend to do better without. We've individually analyzed how each component of groundwater plays an impact on how suitable that area becomes for agriculture. We hope the inferences we have made will not only help substantiate other claims, but can help people better understand how these natural factors play a role in agriculture and how we can possibly reallocate resources more efficiently based on the results given by this analysis.

VI. REFERENCES

[1] A. Gelman, "Missing-data imputation," in *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, 2007, pp. 529-543.

[2] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, pp. 559-572, 1901.

[3] J. MACQUEEN, "SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS," *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.*, vol. 1, pp. 281-297, 1967.

[4] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, "A Density-Based Algorithm for Discovering Clusters," *KDD-96 Proceedings*, pp. 226-231, 1996.

[5] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," *Proceedings of the 20th VLDB Conference* , pp. 487-499, 1994.

GitHub repo:
https://github.com/Rushabh10/CSF415_Project