1 Introduction

1.1 Motivation

Public Health and Safety: Rivers in India are crucial sources of water for drinking, agriculture, and industry. Ensuring the water quality is within safe limits is vital for preventing waterborne diseases and safeguarding public health.

Environmental Impact Assessment: Rivers are integral to the ecosystem. Monitoring their water quality helps in assessing the impact of human activities like industrial discharges, agricultural runoff, and urban development on the aquatic ecosystem.

Climate Change Studies: Understanding the changes in river water quality over time can provide valuable data for climate change research, particularly in assessing the impacts of rising temperatures and changing rainfall patterns on water resources.

1.2 Problem statement:

The project involves the application of clustering techniques to group rivers based on their water quality characteristics. The challenge lies in selecting appropriate features, dealing with data inconsistencies, and interpreting the clustering results to derive meaningful insights into the temporal and spatial variations in Indian river water quality.

1.3 Data description

The data-set consists of 16 features related to water quality measurement, station of river where readings are taken and state. The data is available from year 2012 to 2019.

Each river has different points where sample readings were taken and each such unique point is given a station code. So, a river has multiple station codes.

The quality parameters of river water included in data-set are dissolved oxygen, BOD levels, temperature, potential of hydrogen values, conductivity levels, nitrate nitrite levels, Faecal coliform and coliform count.

2 Data design

The data-set in raw form has 6061 rows, but there were multiple instances of data points of a particular station code for same year. And they were dropped. Also, a new column named Name of river was created to link all the station codes corresponding to a single river.

Filtering data based on Nitrite levels

For the rows where Nitrite levels are more implies more polluted water with increased levels of BOD, reduced levels of oxygen levels and increase in PH. If such variation of these features in those rows with extreme nitrite values are not observed then they are considered as garbage entries and ignored.

For comparison of variation two checks are done, one the same station code is checked with values of other station code of same river for same year. The other check is for same station code values of different years.

After after doing such filtration, only the rows with Maximum nitrate levels less than 150 and minimum nitrate levels less than 15 are kept.

Filtering data based conductivity levels Only two row corresponding to minimum conductivity levels was found not feasible (station code 1326,2360) and were removed.

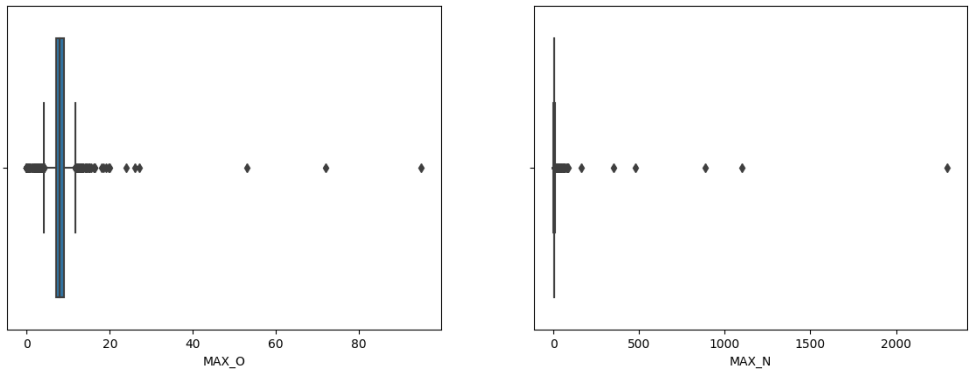

Figure 1: Box plots of DO and nitrites

Filtering data based on dissolved oxygen levels For the station codes where maximum oxygen levels is greater than 40. Imputation is done using mean values replacing the garbage entries.

Filtering data based on hydrogen potential The values of hydrogen potential more than 14 are corrected by using mean value imputation. The station codes 1361 1184 1091 needed correction for maximum hydrogen potential. And station codes 1272 2622 10 needed imputation for minimum hydrogen potential.

Filtering data based on temperature of water Only 1 row of maximum temperature greater than 100 was observed and was removed.

Data filtering based on not mentioned features is not done as they looked feasible values. Box plot of all the features after data processing can be seen from the code.

3 Hierarchical clustering

3.1 Setup

The method used is an agglomerative hierarchical cluster analysis with dissimilarity measure based on Euclidean and Manhattan distances. Scaling is not adopted as it can restrict the influence of extreme values observed in case of some rivers like Yamuna, Ganga etc,.

The features used for clustering are chosen from all after developing a heat-map of covariance matrix. Features used are **minimum oxygen levels, maximum BOD, Maximum hydrogen potential, Minimum conductivity** and **maximum coliform count**. Hierarchical clustering is done after grouping all the rivers by names, and then dendograms are developed for each year separately. Clustering is also performed for all rivers combined and for each year.

Dendograms consisting of rivers **Godavari, Cauvery, Satluj,Krishna, Brahmani, Narmada, Mahanadi,Brahmaputra, Yamuna** and **Ganga** is discussed in this report.

Based on the research papers referred, euclidean distance measure was adopted. Different measures such as ward, complete, average were tried on and all of them gave same hierarchies. Also Manhattan measure with different methods gave similar hierarchies as euclidean measure.

Number of clusters is obtained by manual inspection such that the homogeneous clusters are obtained. For the case of minimum oxygen parameter number of cluster of 66 gave homogeneous plot, cause the number of unique rivers is 306, using only 3 clusters didn't produce homogeneous hierarchies.

3.2 Analysis

All the rivers are classified into 3 clusters, highly polluted, medium polluted and low polluted.For instance, In year 2017, Satluj, Yamuna are placed in cluster 1. Beas, Narmada, Brahmani and Mahanadi are in cluster 2. Godavari, Brahmaputra, Cauvery, Krishna, Ganga and Tungabhadra are in cluster 3.

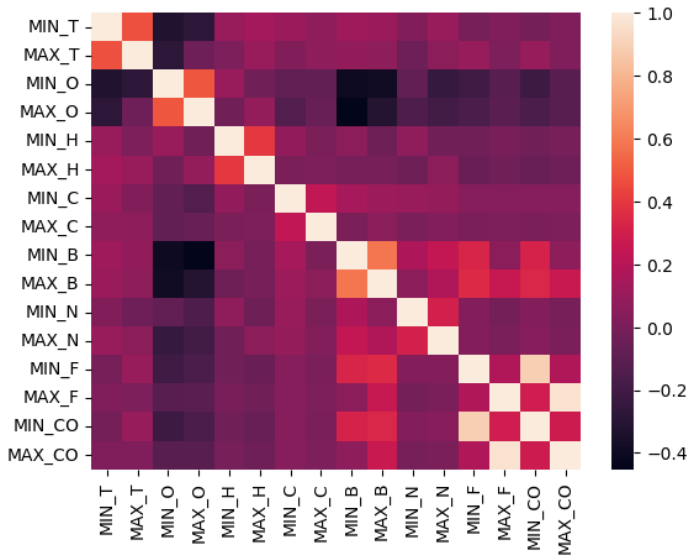

Figure 2: Covariance matrix as heat-map

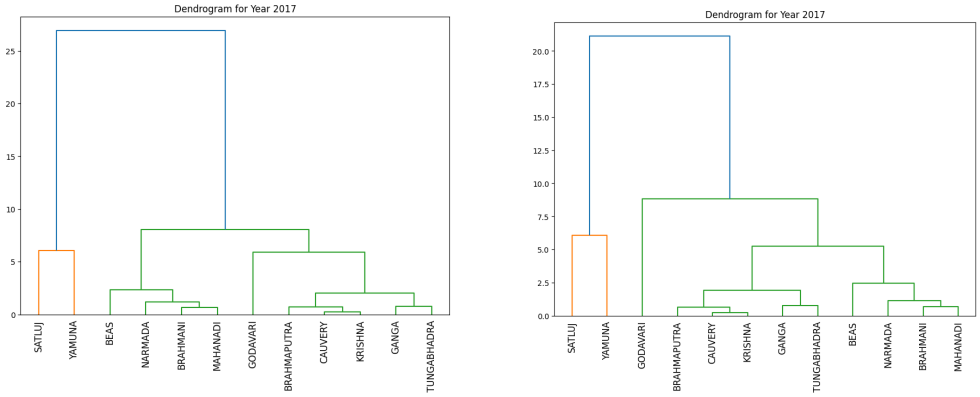

Figure 3: Euclidean with ward method and Manhattan with complete linkage

Ganga was most polluted in 2012 and was placed in single cluster and it improved as years passed, was placed in cluster of fresh water rivers in 2018 but in 2019 it showed more pollution levels and found to be placed in a separate cluster in the model, maybe due to Kumbhmela held in early quarter of year.

Bramhini and Mahanadi are placed in same cluster every year.So are Krishna and Godavari.

4 K-means clustering

4.1 Data processing

Two method to impute mean values are followed. The first method is called Time-based mean imputation and second method is called Spatial-based mean imputation. In the Time-based mean imputation, each station code corresponding to the outliers was checked over the years. If the readings over all the years were consistent, then the mean of the readings were taken and imputed for the outlier values.Figure 4

Spatial-based mean imputation was used when readings of particular station code were not available over the years and when the all the readings over the years was incorrect. In the Spatial-based mean imputation (refer figure 5), the readings for a particular year were taken and the path of the river flow was determined using Google Maps.These were the two main imputation methods incorporated. Other outliers for which mean could not be imputed were removed from the data frame.

In order to obtain a more reliable representation of environmental factors, we employed a methodology that involved calculating average values for key parameters. This included determining the mean of

	Station code	Name of river	srcYear	srcStateName	MIN_T	MAX_T
52	1245	NARMADA	2012	GUJARAT	23.0	26.0
777	1245	NARMADA	2013	GUJARAT	26.0	31.0
1991	1245	NARMADA	2016	GUJARAT	25.0	30.0
2555	1245	NARMADA	2017	GUJARAT	23.0	30.0
3289	1245	NARMADA	2018	GUJARAT	0.0	29.0
4100	1245	NARMADA	2019	GUJARAT	20.0	30.0

	Station code	Name of river	srcYear	srcStateName	MIN_T	MAX_T
52	1245	NARMADA	2012	GUJARAT	23.0	26.0
777	1245	NARMADA	2013	GUJARAT	26.0	31.0
1991	1245	NARMADA	2016	GUJARAT	25.0	30.0
2555	1245	NARMADA	2017	GUJARAT	23.0	30.0
3289	1245	NARMADA	2018	GUJARAT	23.4	29.0
4100	1245	NARMADA	2019	GUJARAT	20.0	30.0

Figure 4: Time based mean imputation

minimum and maximum values, By adopting this approach, the goal was to minimize the impact of potential outliers.

Based on heat map as discussed earlier, it was decided to consider only average temperature, dissolved oxygen, pH, Conductivity level, Nitrate/Nitrite and faecal coliform required for training the K-means clustering model.

	Station code	Name of river	srcYear	srcStateName	MIN_T	MAX_T
2395	1553	YAMUNA	2017	HIMACHAL PRADESH	14.0	24.0
2396	1554	YAMUNA	2017	HIMACHAL PRADESH	14.0	25.0
2397	10004	YAMUNA	2017	HARYANA	0.0	0.0
2398	1119	YAMUNA	2017	HARYANA	0.0	0.0
2399	1120	YAMUNA	2017	DELHI	17.0	33.6
2400	1121	YAMUNA	2017	DELHI	17.3	30.7
2401	1375	YAMUNA	2017	DELHI	18.3	31.2
2402	1812	YAMUNA	2017	DELHI	19.0	33.0

	Station code	Name of river	srcYear	srcStateName	MIN_T	MAX_T
2395	1553	YAMUNA	2017	HIMACHAL PRADESH	14.0	24.0
2396	1554	YAMUNA	2017	HIMACHAL PRADESH	14.0	25.0
1770	1119	YAMUNA	2016	HARYANA	15.35	30.9
2398	1119	YAMUNA	2017	HARYANA	15.50	30.9
2399	1120	YAMUNA	2017	DELHI	17.0	33.6
2400	1121	YAMUNA	2017	DELHI	17.3	30.7
2401	1375	YAMUNA	2017	DELHI	18.3	31.2
2402	1812	YAMUNA	2017	DELHI	19.0	33.0

Figure 5: Spatial based mean imputation

4.2 Clustering and analysis

The K-means clustering analysis was conducted individually for each river, focusing on the averaged parameters derived from six key features over multiple years. Elbow method is adopted to get ideal number of clusters in each case.

The results of the K-means clustering analysis revealed a geographical pattern within the dataset, specifically pertaining to the clustering of river stations based on the states through which the rivers flow. For example, when examining prominent rivers such as the Ganga, the clusters formed align remarkably with the respective states of Uttar Pradesh (UP), Bihar, and West Bengal (WB) that the river traverses. For instance, Cluster 0 predominantly encompasses river stations situated in UP Cluster 1 captures stations within UP and Bihar, Cluster 2 groups together stations in Bihar and WB, and Cluster 3 encapsulates those in West Bengal.

This spatial clustering suggests a clear correlation between the geographic locations of river stations and the identified clusters, emphasizing the influence of regional characteristics on the environmental parameters under consideration. This finding not only underscores the efficacy of the clustering algorithm but also implies potential regional variations in the environmental profiles of the rivers, highlighting the importance of considering geographic context in the assessment of river ecosystems.

Clustering of the river Ganga is described here. There are 4 clusters and each cluster has following properties

Cluster no	avg temp(deg C)	avg DO(mg/L)	avg pH	avg nitrite(mg/L)	avg coliform (MPN/100 ml)
0	21.55	8.41	7.53	0.39	3146.7
1	24.83	7.87	7.97	0.57	20089.44
2	27.25	6.37	7.59	0.74	72231.81
3	26.16	6.49	7.76	1.07	255619.77

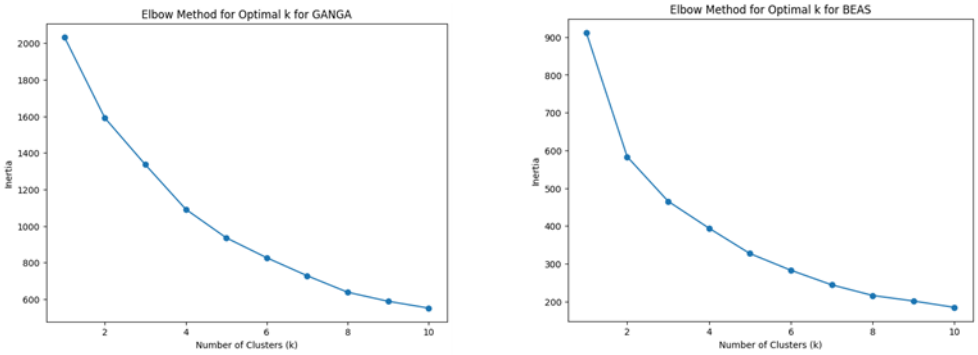

Figure 6: Identifying number of clusters from elbow plots

5 Dimension reduction

PCA facilitated the transformation of the averaged 8-dimensional dataset into a set of principal components, each capturing distinct variance in the data. Notably, the scatter plots revealed discernible patterns, clustering, or dispersion, shedding light on potential correlations or disparities in water quality across different rivers.

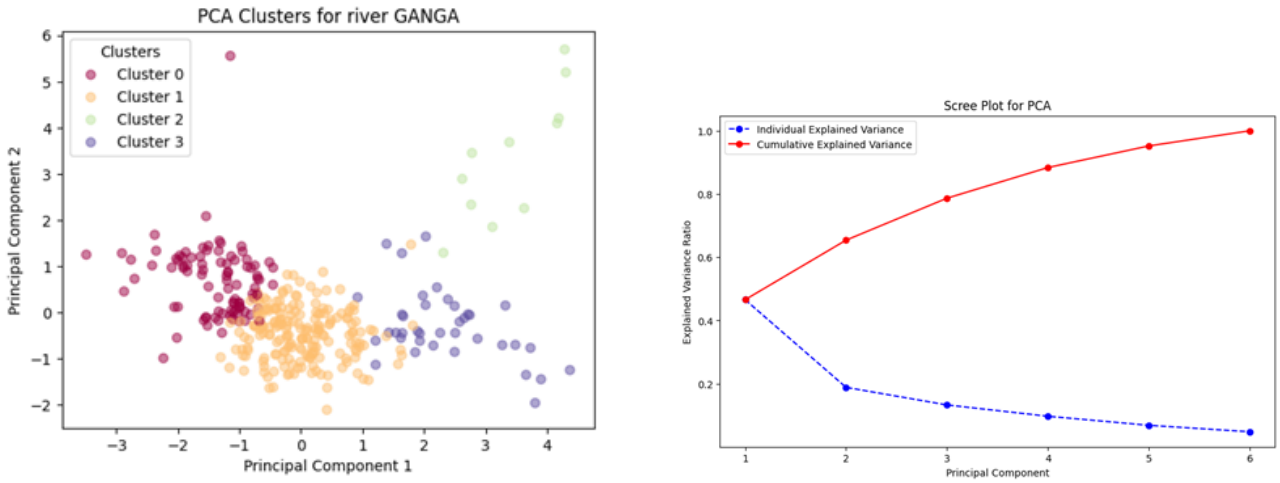

Figure 7: PCA of Ganga and Scree plot of Principal components

The scree plot generated from the Principal Component Analysis (PCA) offers crucial insights into the variance captured by each principal component. In this analysis, the individual explained variances for the first and second principal components, denoted as PCA 1 and PCA 2, respectively, were found to be 0.42 and 0.2. These values signify the proportion of total variance in the data-set that is accounted for by each respective principal component. Notably, PCA 1, with an explained variance of 0.42, emerges as a dominant factor, suggesting that it encapsulates a substantial amount of information inherent in the original data-set. PCA 2, with an explained variance of 0.2, contributes significantly as well, albeit to a slightly lesser extent.

6 Conclusion

The clustering analysis of river pollution data provided valuable insights into the patterns and trends of water quality in Indian rivers. In our analysis of the river pollution dataset, we faced challenges like data incompleteness and variability in measurement standards across different stations, which could be improved with advanced data imputation techniques and standardized measurements. The clustering analysis has significant potential to aid the Indian government and policymakers in identifying pollution

hotspots and formulating targeted environmental policies, thus benefiting society at large. Conclusively, our project not only provided insights into the varying levels of water pollution across regions but also enhanced our understanding of environmental data handling and the practical application of machine learning in real-world scenarios, highlighting the pivotal role of data-driven approaches in environmental conservation and public health.

7 Reference

- **Data set:**Water Quality of Rivers, NDAP Niti Aayog https://ndap.niti.gov.in/dataset/7078
- Water quality monitoring using cluster analysisand linear models: A - Manuela Gonçalvesa and Teresa Alpuim.
- Finding Water Quality Trend Patterns Using Time Series Clustering: A Case Study: Leijun Huang et.al
- Use of Cluster Analysis-A Data mining tool for improved water quality monitoring of River Satluj: Neetu Arora et.al