MA 589 Final Project

Environmental Study

Yumeng Cao, Virushi Patel, Jingyi Li, Liyu Qu

Abstract

This study employs Gaussian Mixture Model (GMM) clustering method to uncover latent patterns in environmental data, which includes four features: average temperature, annual rainfall, air quality index, and forest area percentage. We first used the Principal Component Analysis (PCA) for dimensionality reduction and identifying the most significant features. Then, we performed the 6-component GMM clustering using key features from PC1 and 3-component clustering using key features from PC2. The optimal number of clusters for each model was determined by the Calinski-Harabasz index. Our results provide insights into the underlying structure of this data and may reveal possible information for environmental conservation and monitoring.

Introduction

Understanding the complex interactions between various environmental factors may bring new perspectives to effective conservation and sustainable development. Clustering techniques, such as GMM, provide a powerful approach to discover latent structures within the data. GMMs assume that the data was generated from a mixture of Gaussian distributions, each representing a distinct cluster, which are particularly suitable for this task as it works better for overlapping and non-linearly separable clusters. In this study, we apply GMM clustering to a comprehensive environmental dataset, aiming to uncover hidden patterns since the potential presence of overlapping distributions and the lack of explicit correlations among these features suggested by the initial exploratory analysis. By applying dimensionality reduction techniques like PCA, we identified features that contribute most to data variability. Then, using the key features of the top principal components, we performed two GMM models facilitated by the Expectation-Maximization (EM) algorithm and then explored different numbers of clusters to find the optimal setting. Through this approach, we aim to identify latent relationships between the four available environmental features, which may help people further understand the complex relationships between environmental factors.

Data

1. Data Description

As shown in the following characteristics table, the environmental dataset used in this study consists of 327 observations spanning 8 features (3 categorical features and 5 numerical features), with each feature being complete and having no missing values. From these comprehensive factors, our analysis selected four features, the average temperature, annual

rainfall, air quality index, and forest area percentage, as our primary features which show a more direct connection to environmental conditions.

Table I: Descriptive characteristics of the sample

Feature Names (sample N = 327)	Description - n (%) / mean (sd)
ClimateZone	Arid: 61 (18.7); Continental: 62 (19.0); Polar: 80 (24.5); Temperate: 56(17.1); Tropical: 68(20.7).
PollutionLevel	High: 88(26.9); Low: 83 (25.4); Moderate: 76(23.2); Severe: 80 (24.5).
ConservationStatus	CriticallyEndangered: 80 (24.5); Endangered: 88 (26.9); LeastConcern: 72 (22.0): Vulnerable: 87(26.6).
AverageTemperature	14.69 (14.58)
AnnualRainfall	2112.5 (1116.93)
SpeciesCount	1097.4 (546.94)
AirQualityIndex	256.9 (145.77)
ForestAreaPercentage	51.36 (30.77)

2. Data visualization

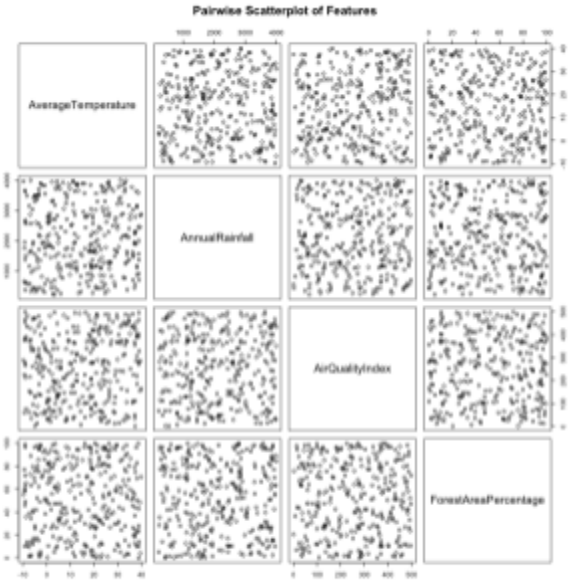

Fig. 1: Pairwise plots of features

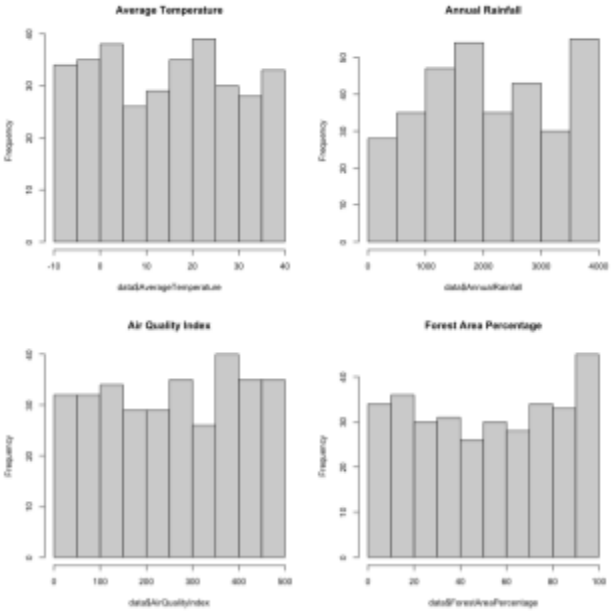

Fig. 2: Distributions of features

Exploratory data analysis was then performed using scatter plots and histograms to gain insights into the distribution and potential subgroups, regarding the pre-selected four features.

The scatter plots showed a lack of clear patterns between feature pairs, with data points widely distributed, suggesting that conventional clustering techniques like K-means may not be suitable for this data. The broad and random distribution of points and absence of apparent correlations indicate potential overlapping subgroups and the presence of multiple underlying Gaussian distributions. The histograms further supported the view of overlapping structures. Both Average Temperature and Annual Rainfall exhibited multi-distributions with multiple peaks. Though Air Quality Index and Forest Area Percentage displayed more uniform distributions, they still show multi-modes, showing the latent structure could not be well represented by singular distribution.

Given the complexity of the underlying patterns revealed within this data, PCA will be used to identify the most important feature combinations. The key features showing in principal components will be then used as input for GMM, implemented through the EM algorithm. GMMs offer a powerful method for modeling the complex, overlapping clusters within the data, allowing for the precise representation of each cluster as specific Gaussian distributions. The insights gained from the exploratory data analysis, combined with the application of PCA and GMM, provide a strong foundation for uncovering the latent patterns within our dataset.

Method

1. **Data Scaling and Principal Component Analysis**

Data scaling was performed using the 'preProcess()' function from the caret package, normalizing the features to have zero mean and unit variance. Following this, we conducted PCA to identify the most important factors, resulting in a selection of 'Air Quality Index' and 'Average Temperature' for PC1 and 'Forest Area Percentage' and 'Annual Rainfall' for PC2. We utilized PCA to reduce the dimensionality of our dataset while preserving the most significant variations in the data, which aids in visualizing and understanding complex relationships.

2. **EM Algorithm for Clustering**

We implemented the EM algorithm to optimally fit a Gaussian mixture model to the normalized data. This step involved initializing the model parameters, iteratively updating these parameters to improve the log-likelihood of the model, and assessing convergence based on a predefined tolerance level. The EM algorithm was chosen because it provides a flexible, probabilistic model for clustering, accommodating the existence of sub-populations within the overall population.

E-Step:

$$Q(\theta|\theta^{(t)}) = \mathbb{E}_{Z|X,\theta^{(t)}}[\log L(\theta; X, Z)]$$

M-Step:

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)})$$

3. Cluster Evaluation

To evaluate the effectiveness of the clustering, we calculated the Calinski-Harabasz index, which provided a measure of the cluster validity. This index helped determine the optimal number of clusters by comparing the variance between clusters to the variance within clusters. The use of the Calinski-Harabasz index enables a more objective assessment of cluster quality, guiding us to choose the number of clusters that maximizes intra-cluster similarity and inter-cluster differences. We used the number of clusters with the highest value to group our data.

Results and Analysis

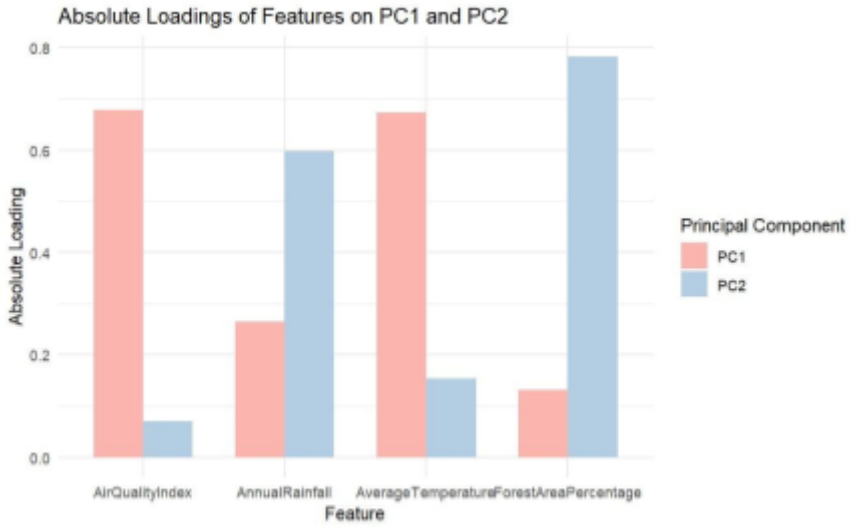

Fig. 3: PCA loadings

Read and Pre-Process The Data:

The features deemed relevant for our investigation, including AverageTemperature, AnnualRainfall, AirQualityIndex, and ForestAreaPercentage, were selected for further analysis. To prepare the data for modeling, preprocessing techniques were applied to standardize the features, ensuring comparability across different variables. PCA was then employed to identify the most influential factors contributing to pollution levels. The PCA results revealed that the first two principal components (PC1 and PC2) captured a significant proportion of the variance in the data, with PC1 explaining 27.67% and PC2 explaining 25.26% of the total variance, respectively. From Figure 3, PCA loadings highlight that PC1 is primarily influenced by Average Temperature and Air Quality Index, as well as PC2 is strongly associated with Forest Area Percentage and Annual Rainfall. Consequently, the researchers perform separate EM analyses for these two groups. Specifically, one EM analysis will focus on the features primarily influencing

PC1: Average Temperature and Air Quality Index. The second EM analysis will target the features associated with PC2–Forest Area Percentage and Annual Rainfall. This approach allows us to develop deeper into the specific groups to uncover unique patterns underneath the dataset.

PC1: Clusters of Average Temperature and AQI

EM Iterations:

The EM algorithm was instrumental in uncovering underlying patterns and optimal cluster structures within the dataset. The EM algorithm was applied to a range of clusters, from 2 to 8, in order to find the optimal number of clusters in the dataset. By systematically exploring a range of cluster sizes from 2 to 8, researchers gain a comprehensive understanding of the dataset's underlying structure. The algorithm was initialized with random initial mu and sigma to avoid convergence to local optima. This could indicate that, despite the increased complexity of having more clusters, the algorithm found a more optimal solution or that the clusters were better separated, leading to faster convergence. Calinski-Harabasz (CH) index, as figure 4 shows, is employed through the process of the EM algorithm to find the appropriate number of clusters. The CH index was utilized as a criterion for evaluating the clustering performance, with higher CH indices indicating better cluster separation. From the line graph, Number 6 has the highest CH index, which implies that dividing the dataset into six clusters provides the most statistically significant differentiation among the clusters. This optimal number of clusters suggests a strong grouping structure within the data, where each cluster represents a distinct group with substantial differences from the others. Thus, we choose 6 clusters for the further analysis. The identification of six clusters through the EM algorithm provides a nuanced understanding of pollution characteristics and spatial variations, enabling targeted environmental interventions. By analyzing the distribution of samples across these clusters, policymakers can identify regions with similar environmental profiles and pollution levels.

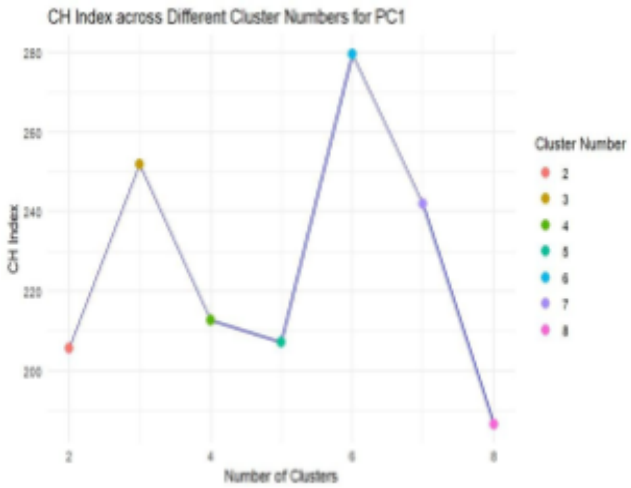

Fig. 4: PC1 CH index

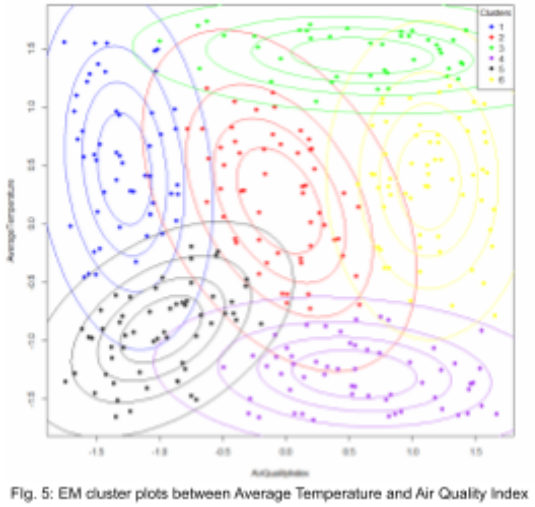
Fig. 5: EM cluster plots between Average Temperature and Air Quality Index

Output Cluster Results:

The output cluster results from the GMM analysis provide valuable insights into the underlying structure of the dataset. The means of the Gaussian distributions represent the centroids of clusters identified by the model. Each mean vector consists of two components: Air Quality Index and Average Temperature. For instance, the mean vectors range from approximately (-1.28, 0.47) to (1.12, 0.37), indicating distinct clusters characterized by different combinations of air quality and temperature. Similarly, the covariance matrices of the Gaussian distributions capture the spread or dispersion of data points within each cluster. Although the specific covariance matrices are not printed in the output, they play a crucial role in determining the shape and orientation of the clusters in the feature space. Upon assigning each sample to the most likely Gaussian distribution based on the calculated responsibilities, the resulting figure (Figure 6) displays the distribution of samples across clusters. The table shows the count of samples assigned to each cluster, ranging from 45 to 59 samples per cluster, with no missing values. This distribution underscores the effectiveness of the GMM analysis in segmenting the dataset into distinct clusters based on the features of air quality and temperature.

Clustering and Denormalization:

The clustering analysis revealed distinct patterns in the dataset, with a total of six clusters identified. The distribution of samples across these clusters varied as Figure 6, with Cluster 5 containing the highest number of samples (59), followed closely by Clusters 4 and 6 with 58 and 57 samples, respectively. Clusters 1, 2, and 3 comprised 52, 56, and 45 samples, respectively. This distribution suggests heterogeneous groupings within the dataset, highlighting the complexity of the underlying data structure. Further investigation into the features contributing to clustering unveiled key insights. These findings underscore the importance of environmental factors, particularly air quality and temperature, in shaping the observed clustering patterns. To contextualize these results, denormalization of the data was performed to restore the original scale of the features. The denormalized values for AQI and Average Temperature were calculated as 248.86 and 17.56, respectively. This transformation facilitates a clearer understanding of the environmental conditions associated with each cluster, aiding in the interpretation of the clustering outcomes.

Table I: De-normalized Cluster Centroids for Top Two Features in PC1

Cluster	AirQualityIndex	AverageTemperature
1	70.39842	21.513557
2	248.85610	17.558304
3	344.47501	35.657016
4	335.26792	-4.089834
5	112.46484	1.511073
6	420.83707	20.116726

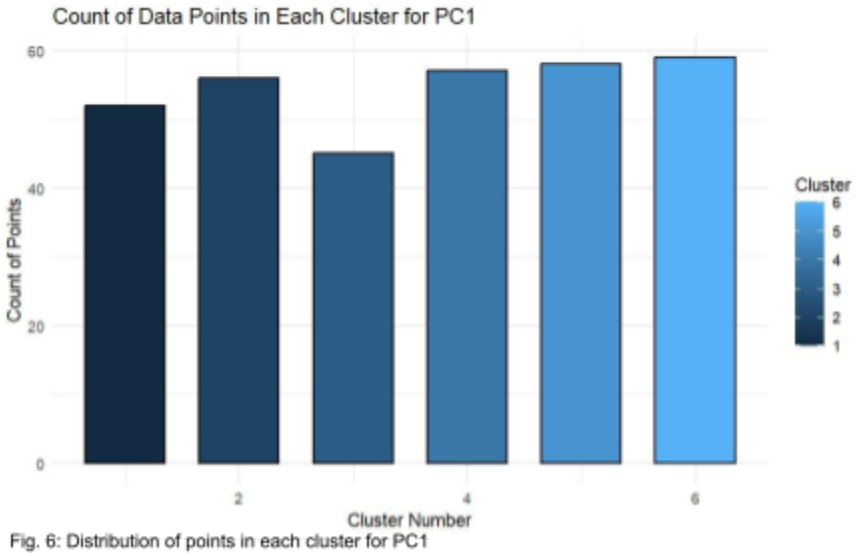

Fig. 6: Distribution of points in each cluster for PC1

Clustering Result for each pair of features:

In our research, we utilized clustering techniques to discern patterns in environmental data, specifically focusing on two key features: Annual Rainfall and Forest Area Percentage. Employing an EM algorithm, we clustered the data into distinct groups to uncover underlying structures and potential correlations between these features. Through visual analysis of the clustering results, we observed compelling patterns and relationships. The plots revealed clusters distinguished by varying shades of blue, red, green, purple, black, and yellow, each representing a unique group of data points. Overlaying ellipses with different levels of significance allowed us to delineate the boundaries of these clusters, offering insight into their distributions and spatial arrangements within the feature space. Our analysis indicated that certain clusters exhibited a higher concentration of data points with similar values for Annual Rainfall and Forest Area Percentage, suggesting a potential association between these variables. Additionally, the

diversity in cluster shapes and sizes hinted at the presence of heterogeneity within the dataset, indicating the existence of distinct subgroups with unique characteristics.

PC2: Clusters of Forest Area Percentage and Annual Rainfall

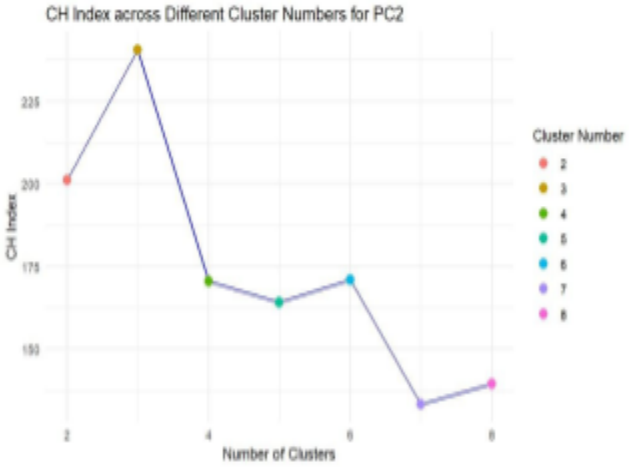

Fig. 7: PC2 CH index

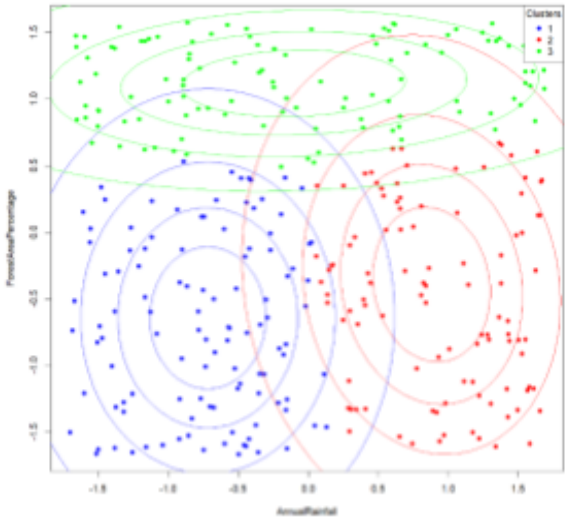

Fig. 8: EM cluster plots between Forest Area Percentage and Annual Rainfall

Final EM Iterations:

The EM algorithm was employed to determine the optimal number of clusters in a dataset using a range of cluster sizes (2 to 8). The results indicated that the best number of clusters, as determined by the CH index, was 3, with a CH index value of 240.74. This suggests that the dataset is best represented by partitioning it into three distinct clusters. Furthermore, when re-evaluating the clustering using only three clusters, the EM algorithm converged after 163 iterations, reinforcing the stability and robustness of the clustering solution. Additionally, Figure 7 shows the CH index varies from 133.31 to 240.74 through cluster numbers from 2 to 8. 3 Clusters get the highest CH index which supports that it is the optimal cluster number for the EM method between Forest Area Percentage and Annual Rainfall.. These results indicate that while the clustering performance improves initially with an increasing number of clusters, it eventually plateaus or even decreases after reaching an optimal cluster size, highlighting the importance of selecting an appropriate number of clusters to avoid overfitting or underfitting the data. Overall, the findings underscore the effectiveness of the EM algorithm in determining the optimal number of clusters and provide valuable insights into the underlying structure of the dataset.

Final Output Cluster Results:

The final parameters obtained from the EM algorithm reveal insightful patterns. The means of the Gaussian distributions represent the central tendencies of the data clusters. For instance, one cluster is characterized by below-average annual rainfall and forest area percentage, indicated by means of approximately -0.72 and -0.64, respectively. Conversely, another cluster exhibits

above-average annual rainfall (mean ≈ 0.88) but relatively lower forest area percentage (mean ≈ -0.39). A third cluster portrays near-average annual rainfall (mean ≈ -0.10) paired with a notably high forest area percentage (mean ≈ 1.12). The covariance matrices provide insights into the dispersion or spread of the data within each cluster, but their detailed interpretation is omitted here for brevity. The mixing weights, denoted as final_lambdas, signify the relative importance of each cluster in the overall dataset. In this analysis, cluster 1 has the highest weight (approximately 0.35), followed closely by clusters 2 and 3 with weights of about 0.33 and 0.32, respectively. Assigning each sample to the most likely Gaussian distribution based on the computed responsibilities allows for a clearer understanding of regional classifications. The resulting table indicates the distribution of samples across the identified clusters. For instance, cluster 1 encompasses 112 samples, while clusters 2 and 3 include 103 and 112 samples, respectively, with no unassigned samples (NA).

Table II: De-normalized Cluster Centroids for Top Two Features in PC2

Cluster	AnnualRainfall	ForestAreaPercentage
1	1310.575	31.54918
2	3095.865	39.26390
3	1997.023	85.72126

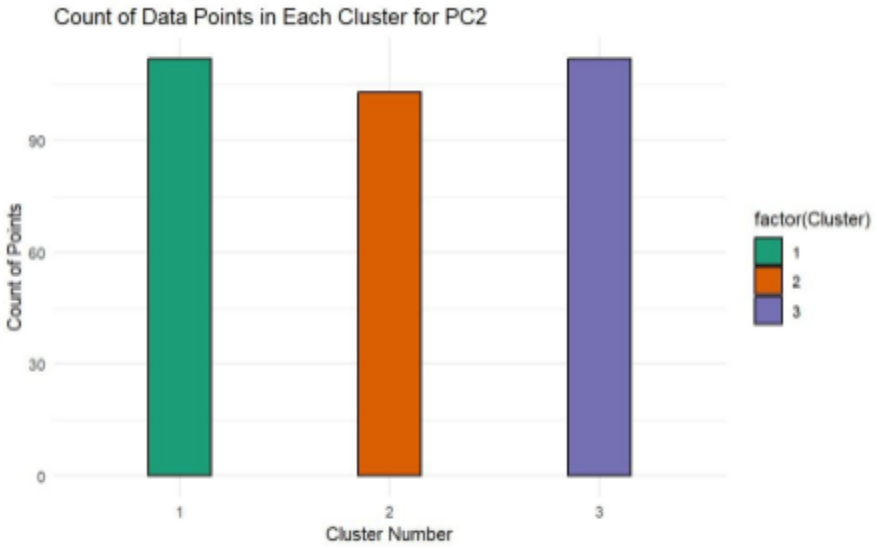

Fig. 9: Distribution of points for each cluster in PC2

Final Clustering Result for each pair of features:

After calculating the Calinski-Harabasz indexes, we determine the number of clusters with the highest value as the final result. We then generate plots where the clusters are clearly divided, and their centers are shown. Regions are grouped into clusters using the features of Average

Temperature and AQI, while the features of Area Percentage and Annual Rainfall result in regions being grouped into 3 clusters. Furthermore, the identification of potential hotspots, such as clusters with consistently high pollution levels across multiple environmental factors, enables proactive measures to mitigate pollution and protect vulnerable populations. By focusing on areas with the greatest environmental challenges, policymakers can implement targeted policies, infrastructure improvements, and community initiatives to reduce pollution levels and improve overall environmental quality.

Conclusion:

In conclusion, this research project leveraged GMM clustering alongside PCA to reveal latent patterns within environmental data encompassing features including average temperature, annual rainfall, air quality index, and forest area percentage. Through further analysis and iterative modeling, we uncovered underlying relationships and distinct clusters within the dataset, shedding light on the complex interplay between environmental factors. Our findings revealed the presence of distinct clusters, each characterized by unique combinations of environmental attributes. By employing sophisticated clustering techniques and dimensionality reduction methods, we successfully capture the complexity of overlapping and non-linearly separable data. By applying EM algorithm and rigorous evaluation like Calinski-Harabasz index, we arrived at robust clustering solutions that offer insights for environmental conservation and monitoring. The identification of six clusters for the features of average temperature and air quality index, and three clusters for forest area percentage and annual rainfall, underscores the nuanced variability within the data. These clusters provide valuable insights into regional environmental profiles, enabling targeted conservation efforts and pollution mitigation strategies. Moving forward, the insights gathered from this research could contribute to environmental policy and resource allocation. For instance, the cluster characterized by high AQI and low Average Temperature may indicate urban areas with industrial emissions or vehicular pollution, while clusters with high Annual Rainfall and Forest Area Percentage could represent rural regions prone to agricultural runoff or deforestation-related pollution. These insights allow decision-makers to prioritize resources and interventions based on the specific needs of each cluster. By understanding the spatial distribution of environmental factors and pollution levels, policymakers could prioritize interventions and allocate resources effectively, ultimately fostering sustainable development and safeguarding ecological health for future generations.