MA 589 Final Project
Environmental Study

Yumeng Cao, Virushi Patel, Jingyi Li, Liyu Qu

Introduction

Goal: Identifying latent relationships among key environmental features through Gaussian Mixture Models (GMM), which may reveal insights into environmental conservation.

Why clustering: Clustering could help to reveal hidden structures and relationships in data, which would be helpful for multidimensional environmental datasets.

Why GMM: Suitable for data with overlapping and non-linearly separable clusters, as it assumes data is generated from multiple Gaussian distributions, each representing a distinct environmental pattern.

Study Approach:

1. Principal Component Analysis (PCA) for dimensionality reduction and identifying the most significant features
2. (1) 6-component GMM clustering using Key Features in PC1
   - Identify the optimal number of clusters using Calinski-Harabasz Index (n = 6)
   - EM Iterations for Optimal Cluster Number
3. (2) 3-component GMM clustering using Key Features in PC2
   - Identify the optimal number of clusters using Calinski-Harabasz Index (n = 3)
   - EM Iterations for Optimal Cluster Number

Data

The dataset consists of 327 observations and 8 features (3 categorical features and 5 numerical features), having no missing values.

To simplify the problem, 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.

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)**

Table I: Descriptive characteristics of the sample

Data

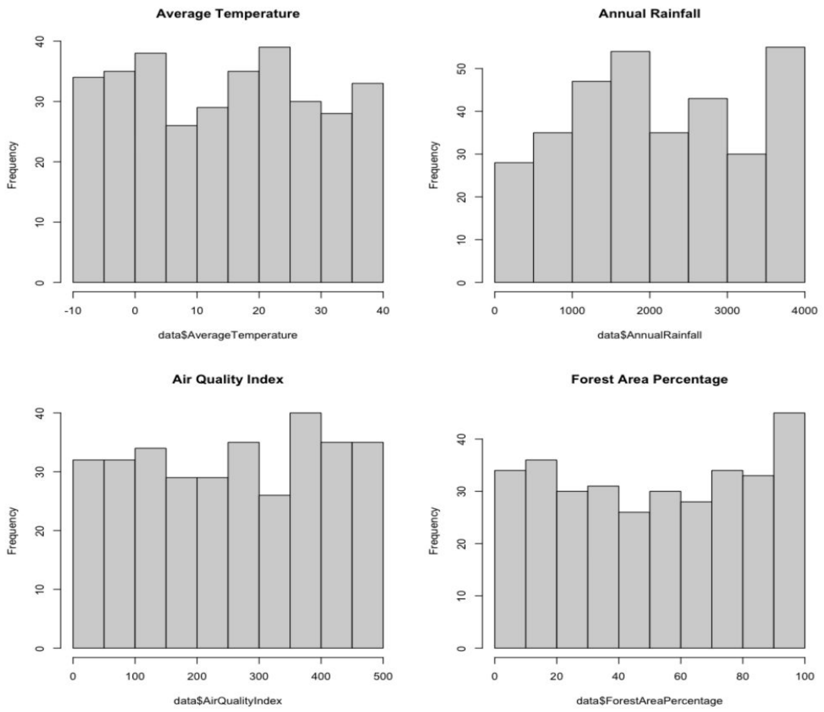

Fig.1: Distribution of Features

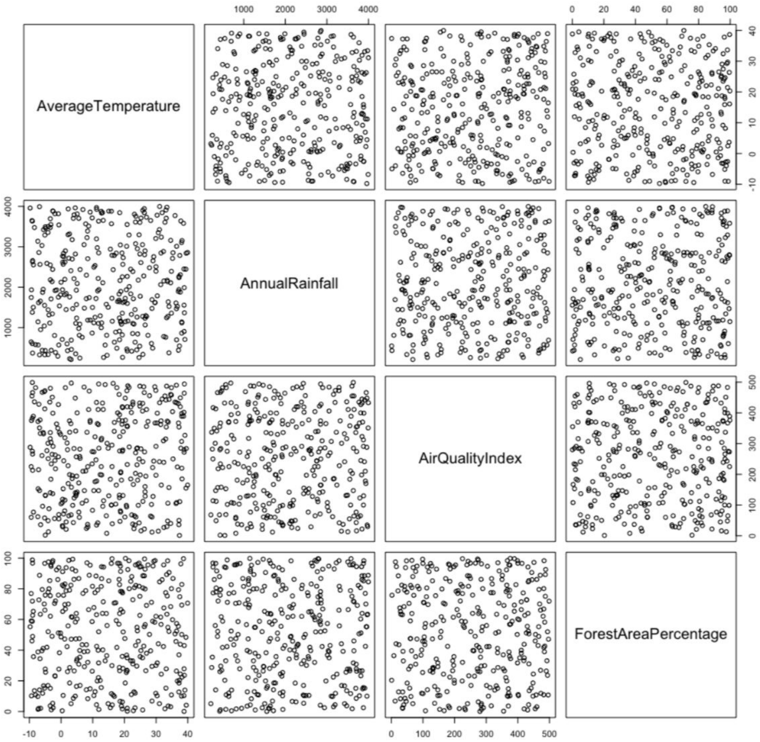

Fig.2: Pairwise Plot of Features

Method

- Conduct PCA to identify the most important factors for PC1 and PC2

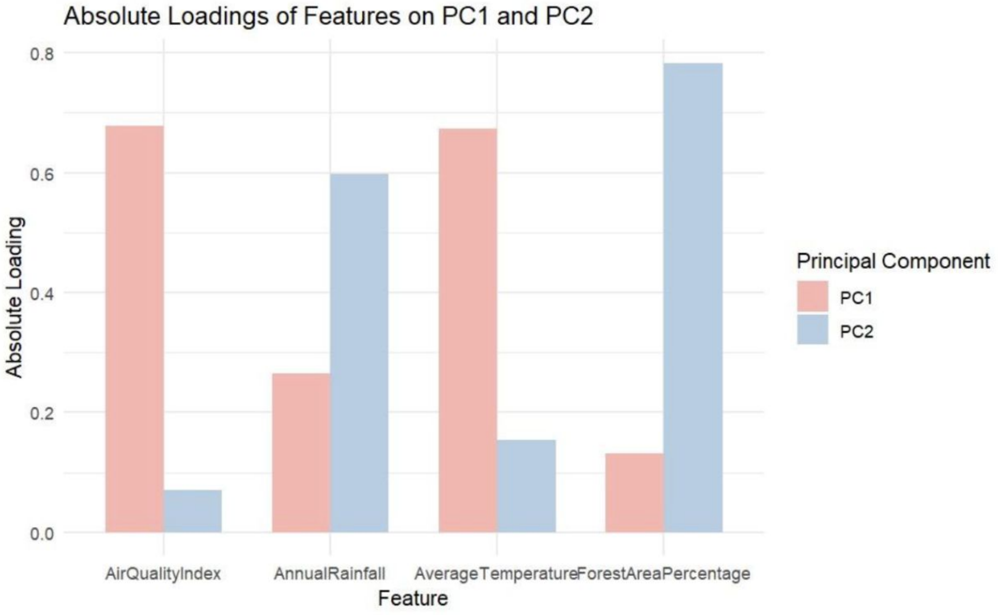

Fig. 3: PCA loadings

Method

- Use EM algorithm to find the appropriate number of clusters by calculating the Calinski-Harabasz index

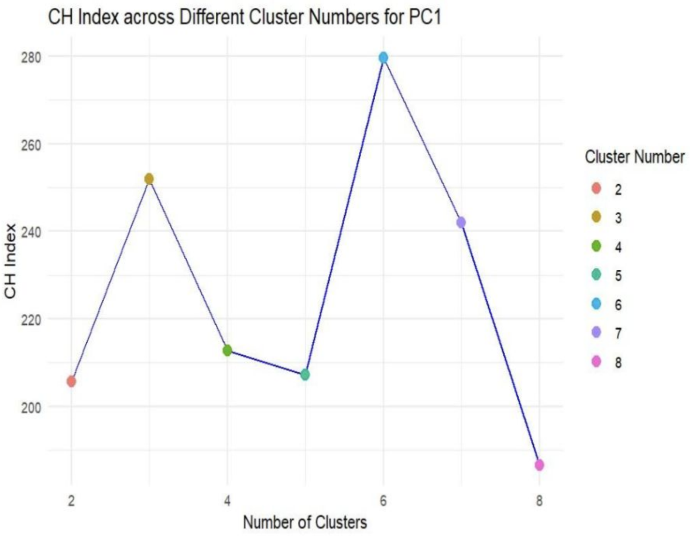

Fig. 4: PC1 CH index

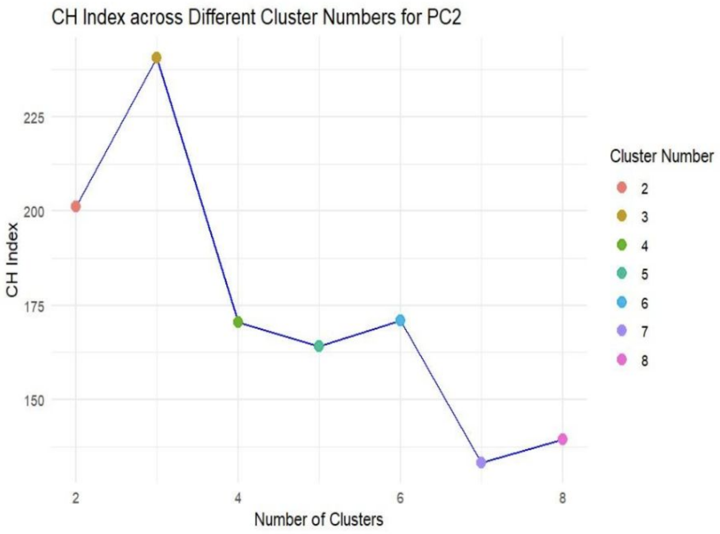

Fig. 7: PC2 CH index

Method

- Use prof's em update function
- Calculate $E[Z_i|X_i;\theta_i]$ by softmax function instead of sigmoid function.

```
em_update <- function (x, em) {
  # [ E-step ]
  ll <- qlogis(em$lambda) # logit(lambda)
  f1 <- apply(x, 1, ldmvnorm, em$mu1, chol(em$sigma1))
  f2 <- apply(x, 1, ldmvnorm, em$mu2, chol(em$sigma2))
  p <- plogis(f1 - f2 + ll) # E[Z_i | X_i; theta_t]
  # Q(theta_t; theta_{t-1}), just for information:
  Q <- sum(p * (f1 + log(em$lambda)) + (1 - p) * (f2 + log(1 - em$lambda)))
  # [ M-step ]
  lambda <- mean(p)
  mu1 <- apply(x, 2, weighted.mean, p)
  mu2 <- apply(x, 2, weighted.mean, 1 - p)
  sigma1 <- cov.wt(x, wt = p, center = mu1, method = "ML")$cov
  sigma2 <- cov.wt(x, wt = 1 - p, center = mu2, method = "ML")$cov
  list(lambda = lambda, mu1 = mu1, mu2 = mu2,
       sigma1 = sigma1, sigma2 = sigma2,
       p = p, Q = Q)
}
```

Result – PC1

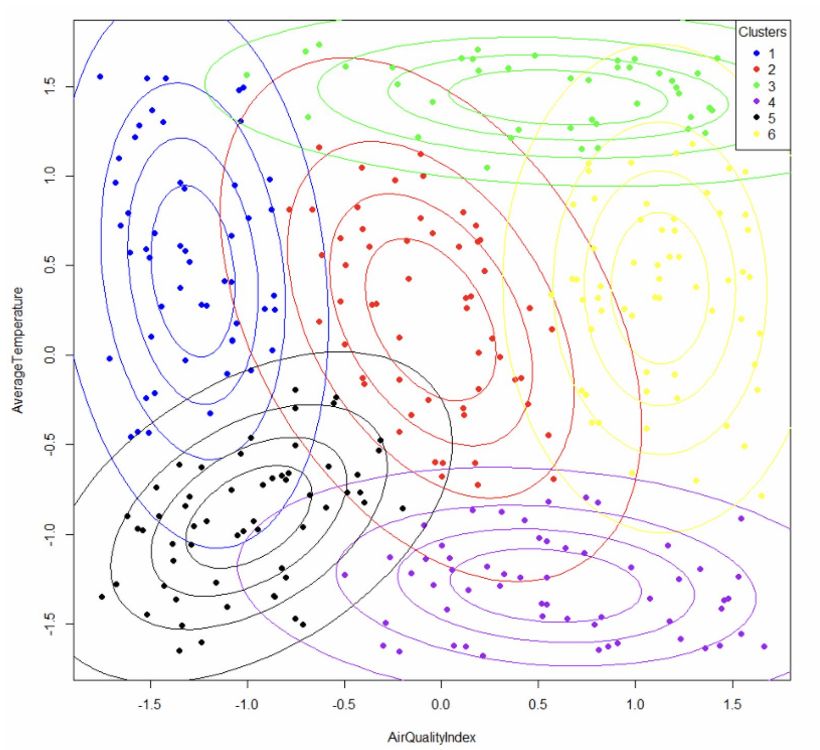

FIg. 5: EM cluster plots between Average Temperature and Air Quality Index

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

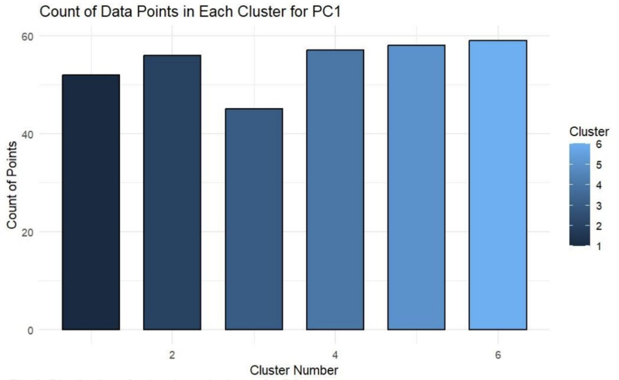

Fig. 6: Distribution of points in each cluster for PC1

Result – PC2

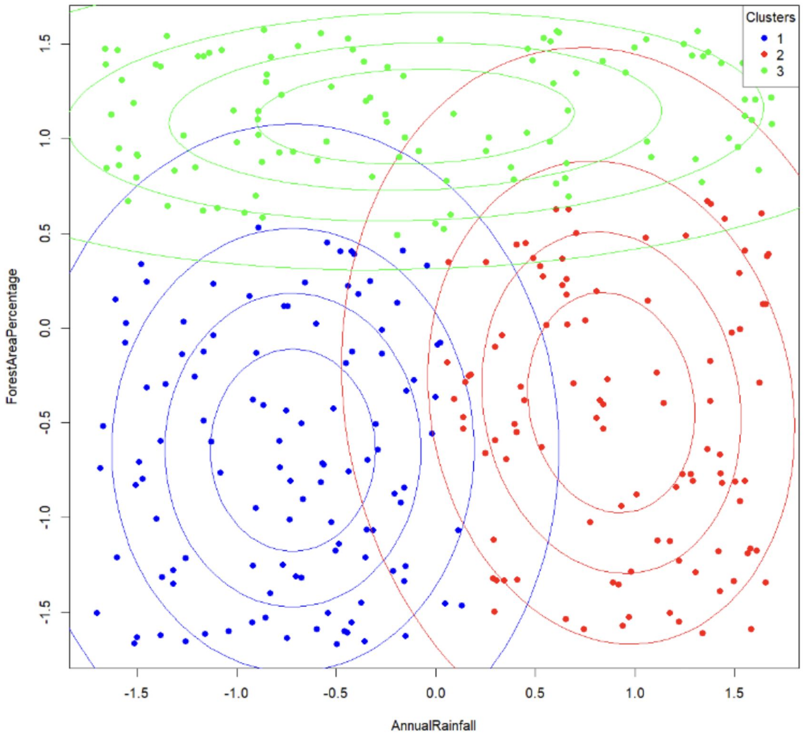

FIg. 8: EM cluster plots between Forest Area Percentage and Annual Rainfall

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

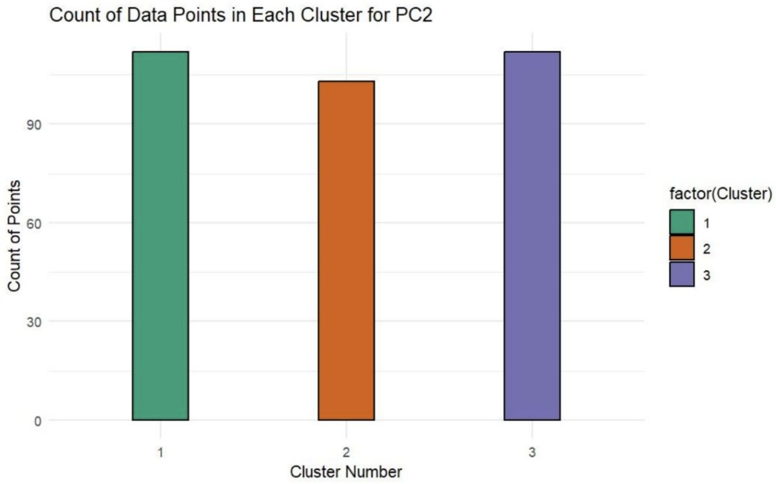

Fig. 9: Distribution of points for each cluster in PC2

Conclusion

- The identification of multiple clusters across different environmental features highlights the variability within the dataset.
- The 6-Component GMM for Temperature and AQI identified 6 clusters showing varying combinations of air quality and temperature.
- The 3-Component GMM for Rainfall and Forest Area identified 3 clusters, each indicating different correlations between forest area and rainfall.
- The results of the two models demonstrates its ability to capture complex latent patterns within environmental factors that are not linearly separable, and also provided insights into regional environmental characteristics that may contributes to environmental conservation.
- The insights gathered hold potential contributions to environmental policy and resource allocation, prioritizing interventions based on the specific needs of each clus.