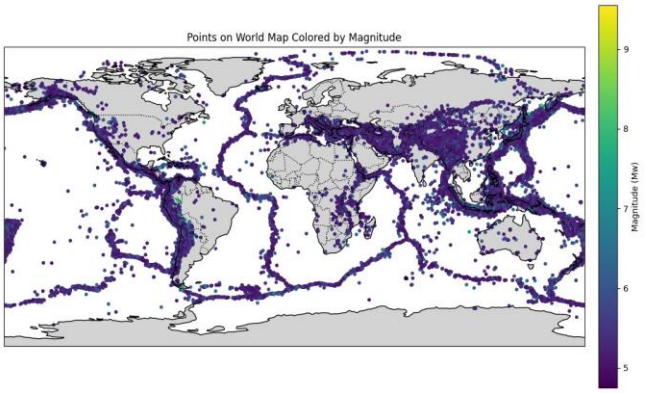
# Clustering of Earthquake Points

**Pravin Ravi (CE22B092)**

## Introduction

Understanding earthquake patterns is crucial for seismic risk assessment, disaster preparedness, and scientific research. Clustering techniques help identify seismic activity patterns by grouping earthquakes based on their geographical and depth-related attributes. This report explores various clustering methods, including density-based, partition-based, fuzzy, and model-based approaches, and evaluates their effectiveness in analyzing earthquake data.

## Data Represented in Map (After Processing)




## Clustering Techniques

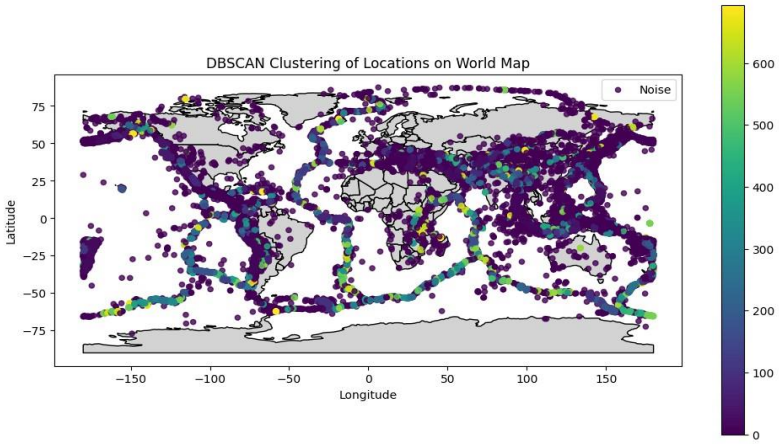
Several clustering techniques were applied to analyze earthquake data, each with distinct strengths and limitations.

### 1. Density-Based Clustering

#### DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

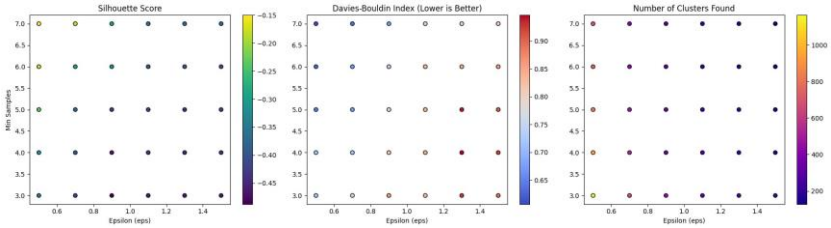
**Parameters Used:**

- `eps=0.5`: Defines the radius within which neighboring points are considered part of a cluster.
- `min_samples=10`: Minimum number of points required to form a cluster.
- `metric='haversine'`: Used to account for spherical distance between geographic coordinates.




**Findings:**

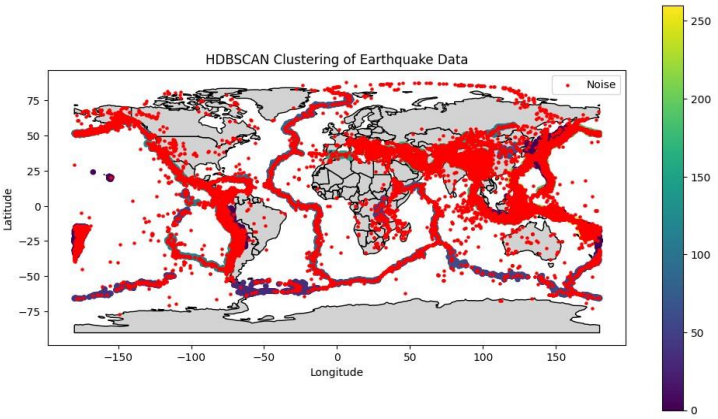
- DBSCAN effectively identified dense seismic clusters while classifying sparse events as noise.
- It struggled with variable-density clusters, leading to fragmentation in some regions.
- The `eps` parameter required tuning based on dataset characteristics to optimize performance.

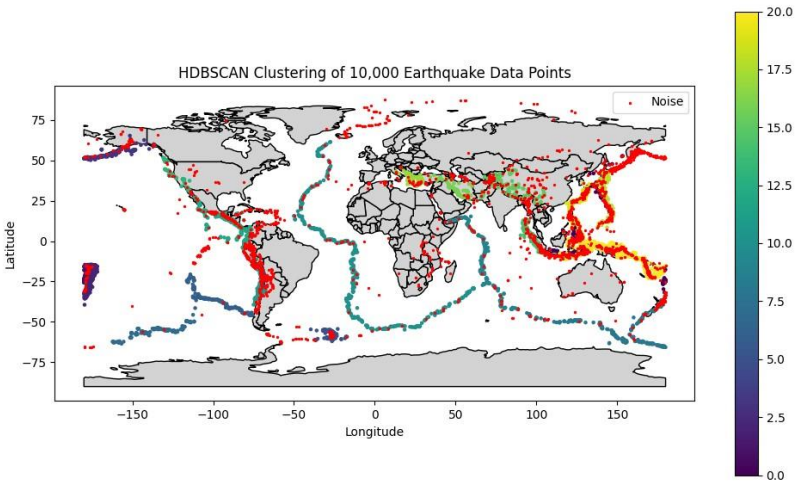



**HDBSCAN (Hierarchical DBSCAN) Parameters**

**Used:**

- `min_cluster_size=50`: Minimum number of points required to form a cluster.
- `min_samples=10`: Controls the level of noise filtering.
- `metric='haversine'`: Used for distance computation in geospatial data.



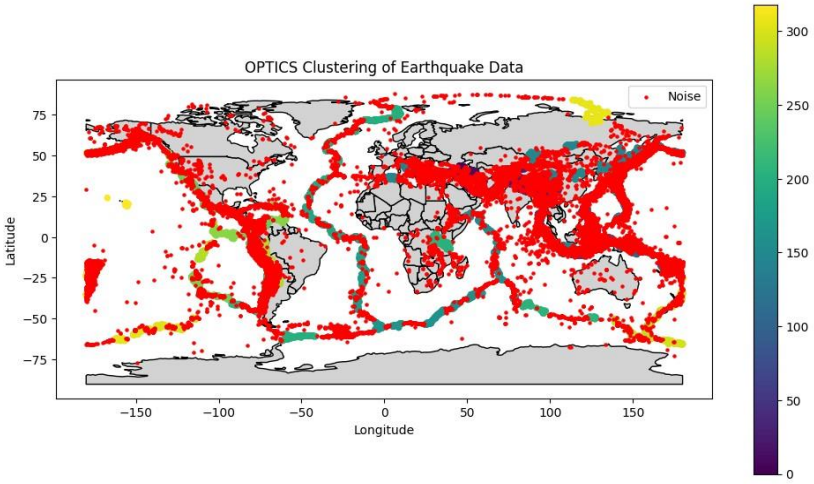

**Findings:**

- HDBSCAN performed better than DBSCAN by handling variable-density clusters more effectively.
- It automatically determined the number of clusters, removing the need to tune `eps` manually.
- HDBSCAN provided a **probabilistic cluster membership**, making it useful for uncertain seismic zones.

**OPTICS (Ordering Points to Identify the Clustering Structure) Parameters**

**Used:**

- `min_samples=10`: Minimum number of points required to form a cluster.
- `xi=0.05`: Minimum steepness of density variation for cluster detection.
- `min_cluster_size=50`: Minimum size of a valid cluster.
- `metric='euclidean'`: Euclidean distance used for clustering.




**Findings:**

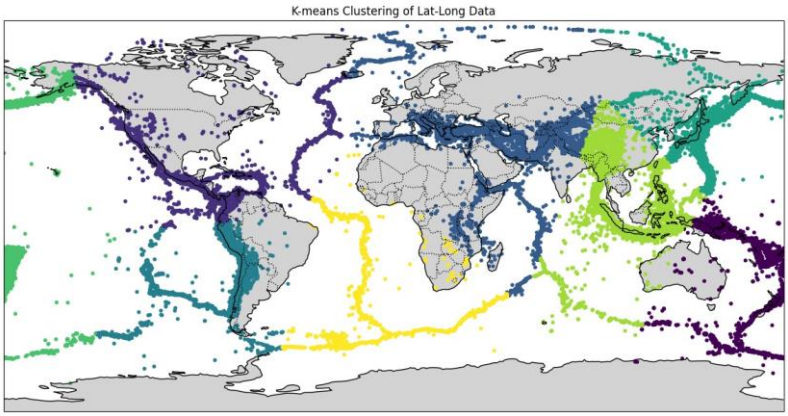
- OPTICS effectively identified clusters in denser seismic regions.
- Many points were classified as noise due to density variations, which is a known limitation of density-based clustering.
- It performed well for moderate datasets (~10,000 points) but struggled with large datasets (>70,000 points).

## 2. Partition-Based Clustering

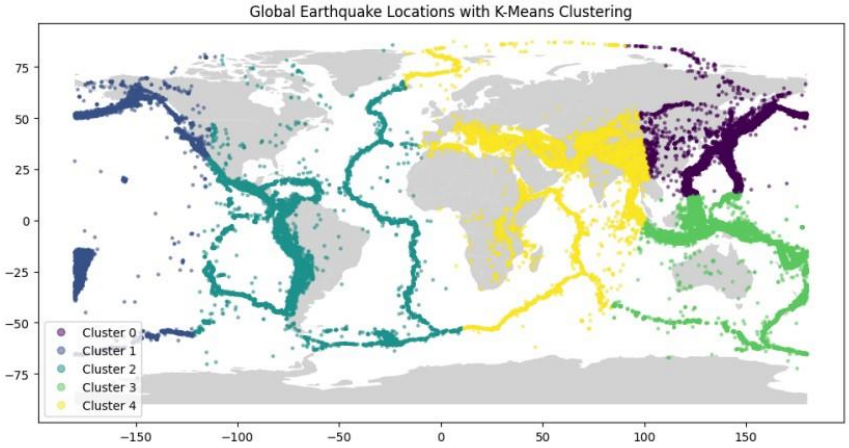
### K-Means Clustering Parameters

**Used:**

- `k=8/5`: Chosen as the cluster number.
- `random_state=0`: Ensured reproducibility.



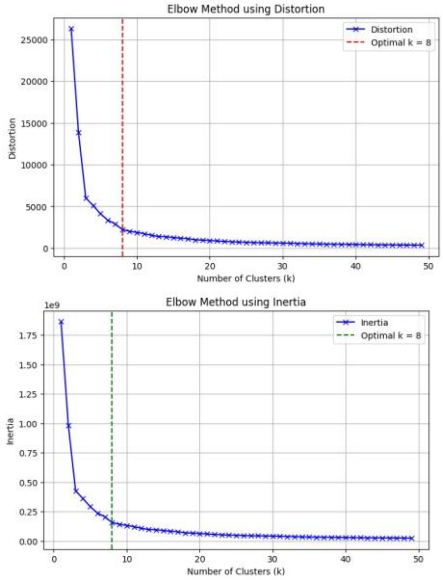
Features: Lat, Lon, Depth (8 clusters)



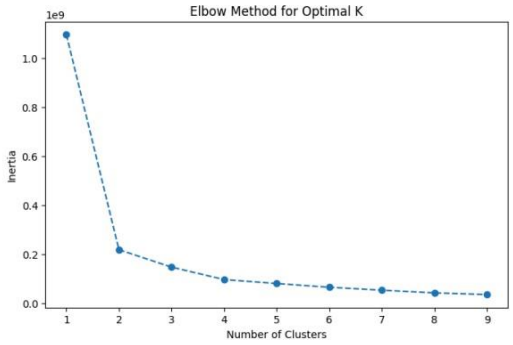

Features: Lat, Lon (5 clusters)

**Findings:**

- K-Means provided well-defined cluster centers.
- However, it assumes clusters are spherical, which may not align well with real-world earthquake distributions.
- The optimal number of clusters was determined using the **Elbow Method** refining the value of `k`.




Features: Lat, Lon, Depth (8 clusters)



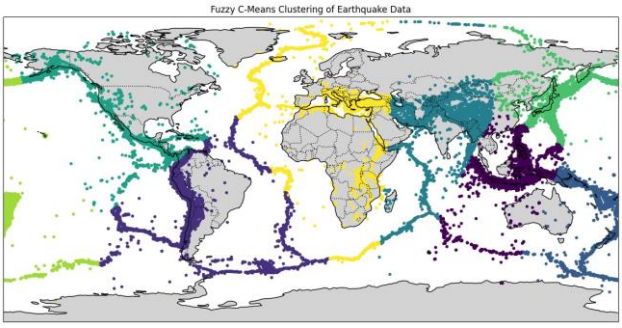

Features: Lat, Lon (5 clusters)

## 3. Fuzzy Clustering

### Fuzzy C-Means (FCM) Parameters

**Used:**

- `k=8`: Number of clusters.
- `m=2`: Fuzziness coefficient, allowing data points to belong to multiple clusters.
- `error=0.005`: Convergence threshold.
- `maxiter=1000`: Maximum iterations for stability.




**Findings:**

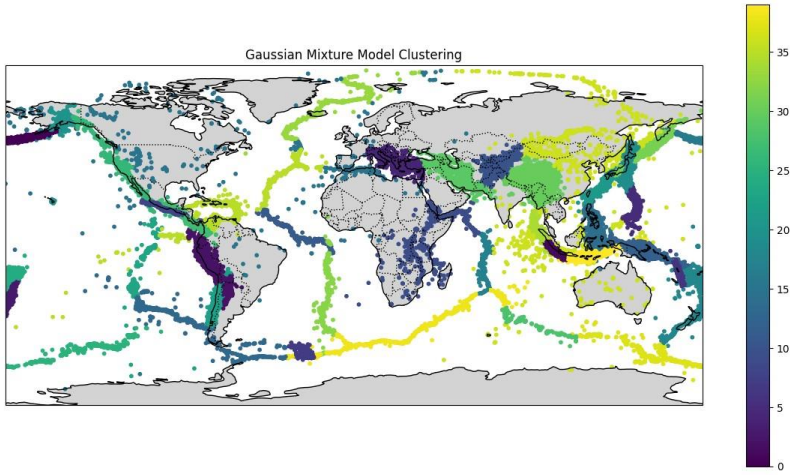
- FCM provided probabilistic cluster memberships, improving handling of boundary cases.
- It required more computational resources and performed better for overlapping seismic zones.

## 4. Model-Based Clustering

### Gaussian Mixture Model (GMM)

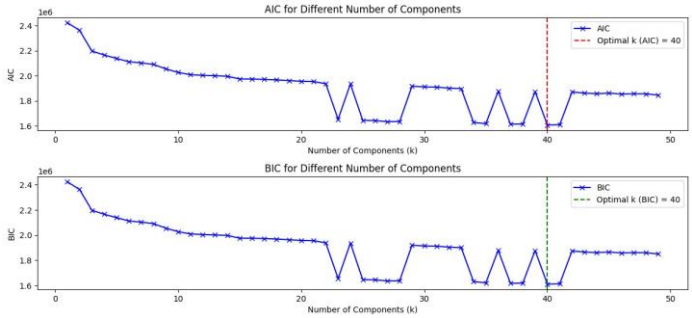
**Parameters Used:**

- `n_components=40`: Initial number of clusters.
- `covariance_type='full'`: Allowed flexibility in cluster shapes.
- `random_state=42`: Ensured reproducibility.




**Findings:**

- GMM performed well for large datasets by modeling elliptical clusters.
- **Bayesian Information Criterion (BIC)** and **Akaike Information Criterion (AIC)** were used to determine the optimal cluster number, minimizing overfitting.
- The BIC gradient method provided a refined estimate of the optimal cluster count.




# Cluster Validation

To assess clustering quality, the following metrics were used:

1. **Elbow Method (Distortion & Inertia)**: Determines the optimal `k` in K-Means.
2. **Silhouette Score**: Evaluated cluster compactness and separation.
3. **BIC & AIC Scores**: Used for GMM optimization.
4. **Cluster Stability Index**: Measured consistency of clusters across different runs.

# Conclusion

- **Density-based methods** like DBSCAN, HDBSCAN, and OPTICS does not work well for seismic clustering as it struggles with sparse data.
- **Partitioning methods** such as K-Means are efficient but assume clusters are spherical, which may not reflect earthquake distributions.
- **Fuzzy clustering** captures uncertainty better but requires significant computational power.
- **Model-based clustering (GMM)** provides flexibility in cluster shapes and performs best for large datasets.