Machine learning for clustering

Damien Chanet
ENSIIE, January

Introduction

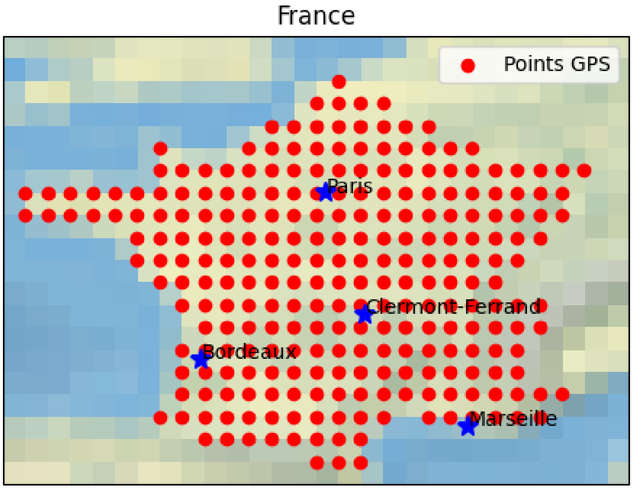
The aim of this project is to perform a segmentation of the French territory based on Temperature and Wind time series gathered at n = 259 grid points using different clustering methods.

The datasets

In this machine learning project, we have three datasets :
- dataTemp.csv provides the temporal evolution of the temperature at an hourly sampling rate for a given year (p = 8760 hours) for 259 grid points,
- dataWind.csv provides the temporal evolution of the wind speed at an hourly sampling rate for a given year (p = 8760 hours) for 259 grid points,
- dataGPS.csv provides the GPS positions (longitude and latitude) of the 259 time series grid points. The grid points are located as follow on the French territory

Data observation

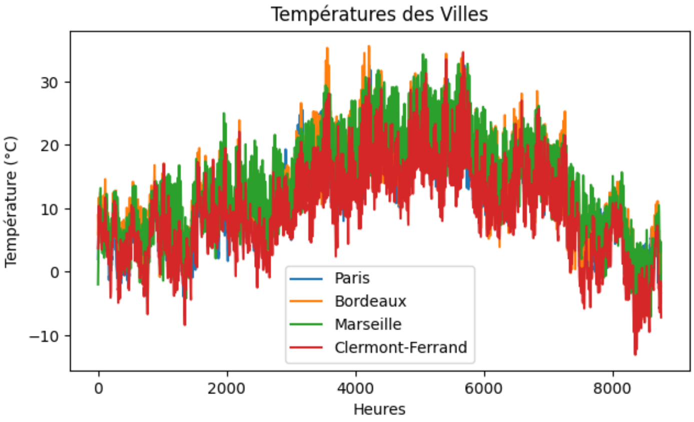
Before clustering the data, we need to understand and visualize the data. Therefore, we chose 4 different french cities : Paris (Latitude : 48.5 Longitude : 2.2), Bordeaux (Latitude : 44.8 Longitude : -0.57), Clermont-Ferrand (Latitude : 45.8 Longitude : 3.1) and Marseille (Latitude : 43.3 Longitude : 5.4).

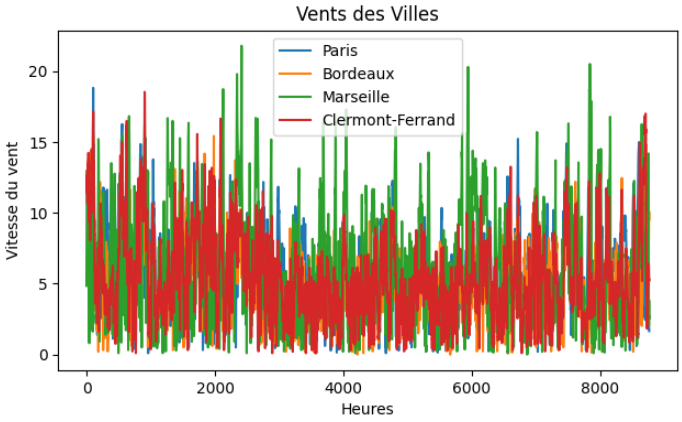


We need to find the 4 grid points of the GPS dataset which have the closest GPS position to the four cities. In order to get those grid points, we create the function find_closest_point which returns the index of the closest grid point of a position.
- Closest grid point index for Bordeaux : 32 - TEMP4439
- Closest grid point index for Bordeaux : 182 - TEMP4439

- Closest grid point index for Marseille : 244 - TEMP4707
- Closest grid point index for Clermont-Ferrand : 156 - TEMP4277

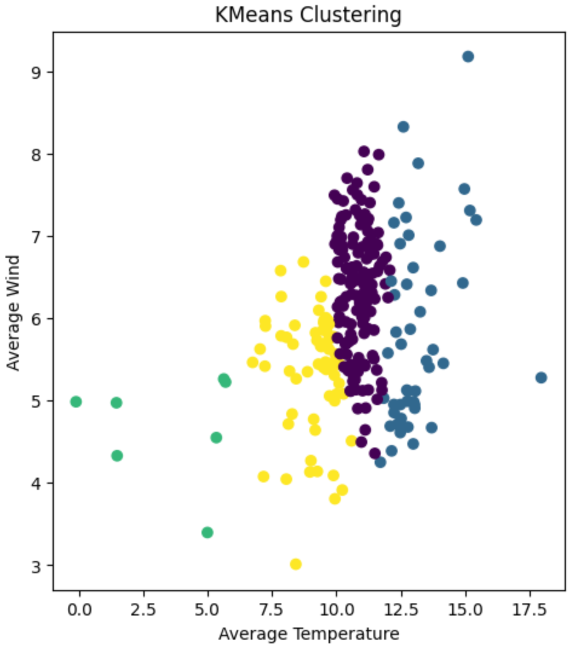
After that, we plot the wind and the temperature data of the four grid points and we have the two graphics below.

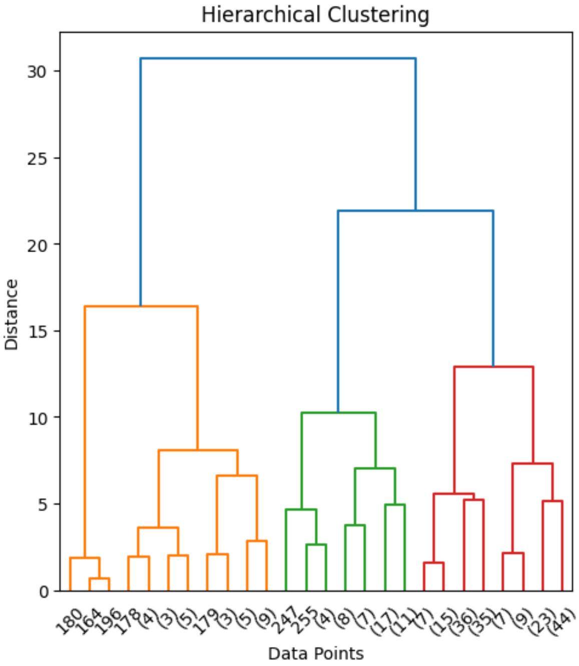




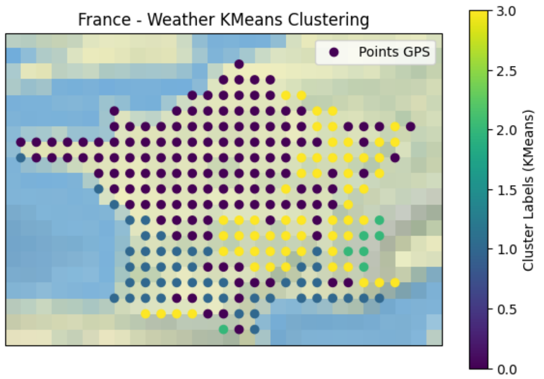
I) Preliminary

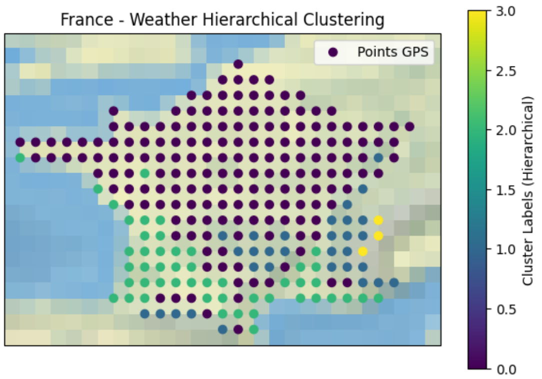
Before entering in details, we try a naive segmentation of the weather. We compute the average temperature and wind speed at each grid point. Then, we perform KMeans clustering with 4 clusters, and hierarchical clustering with 4 clusters. The results are then displayed and a dendrogram for hierarchical clustering is plotted.





We can see how well the data points are grouped together in distinct clusters but it doesn't represent the reality of our weather.



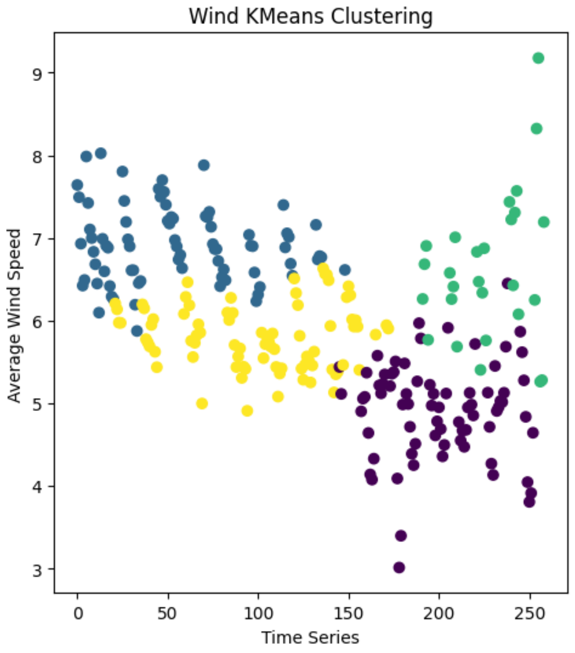

France - Weather Hierarchical Clustering
Points GPS
Cluster Labels (Hierarchical)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

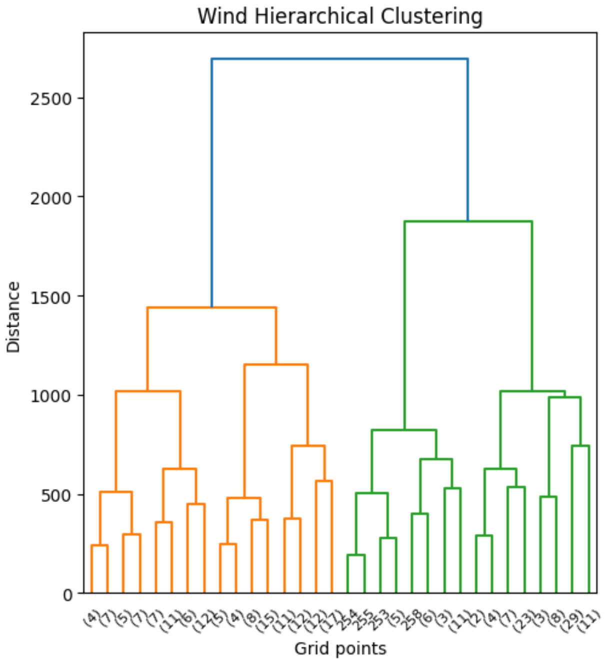
Here are the grid points on the map of France. We see differences in the north-east of France between clusters of Kmeans and Hierarchical clustering but there is some similarity in the south of France.

II) Wind clustering

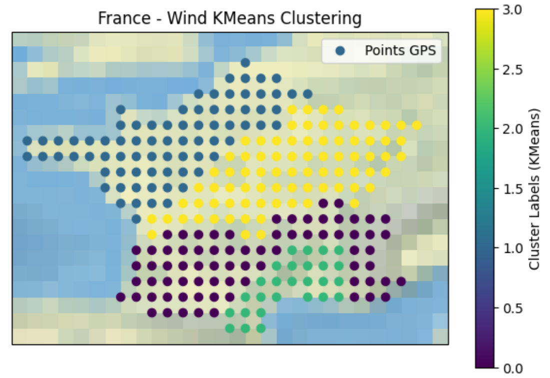
A) Raw data

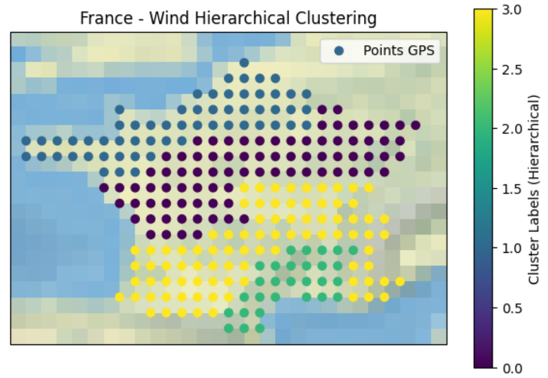
We study and compare the use of the Kmeans and the hierarchical clustering to provide a segmentation into 4 groups of the Wind using the raw time series. For each of the 259 points, the Kmeans algorithm examines the entire time series (p = 8760 values for each point in our case) and assigns a cluster label (0 to 3) based on the similarity of the entire time series. The resulting array kmeans_labels will have 259 values, each indicating the cluster assignment for the corresponding point in the 'wind' dataset.





We can see clear separations between the vertical lines, and the branches are well-defined, it indicates that the algorithm has identified distinct clusters in your data.




France - Wind Hierarchical Clustering
Points GPS
Cluster Labels (Hierarchical)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

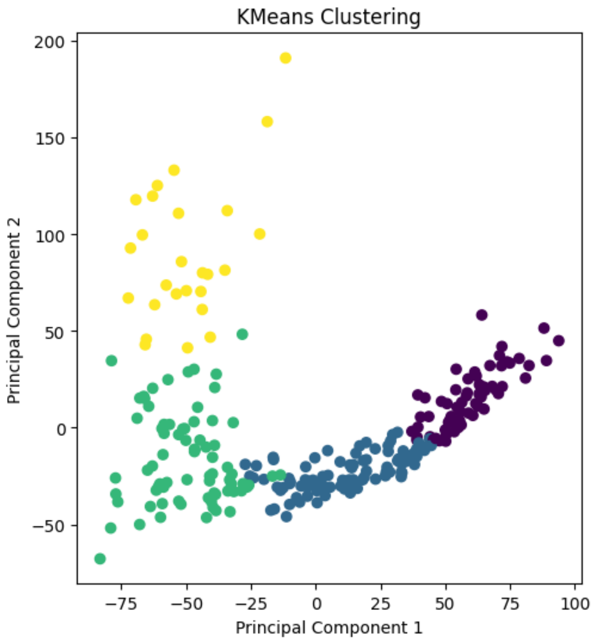
B) Feature extraction

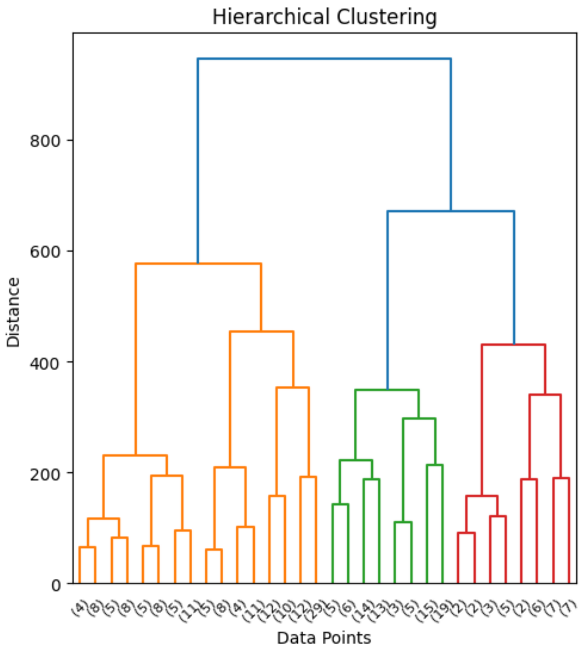
Functional Principal Component Analysis (fPCA) can be used to analyze and reduce the dimensionality of time series data.

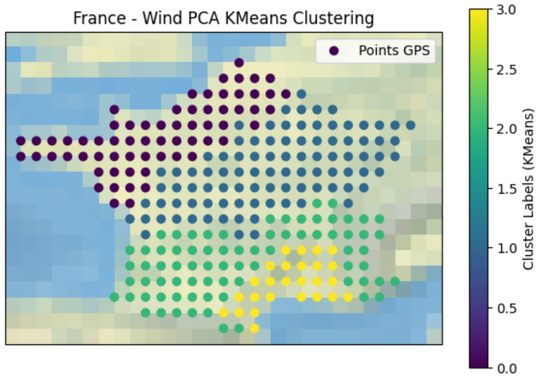
We need to standardize the wind data to ensure that each time series has a mean of 0 and a variance of 1 because this step is crucial for PCA. Then PCA() is applied to the standardized wind data. The explained_variance_ratio_ attribute provides the variance explained by each principal component. The cumulative variance explained is computed by taking the cumulative sum of the explained variance ratios. We choose a threshold (in our case =0.95) for cumulative variance explained (95%). The code then determines the number of components needed to exceed or reach this threshold. Finally, the wind data is transformed using the selected number of principal components. The output indicates that we need 47 components to achieve this threshold. That means 95% of the data are explained by 47 components. We need to keep 47 components because this value is the balance between compression and information retention. While it's important to retain enough components to capture most of the variance, keeping too many components may not be efficient and might not offer substantial benefits. We want to strike a balance between reducing dimensionality and retaining meaningful information.

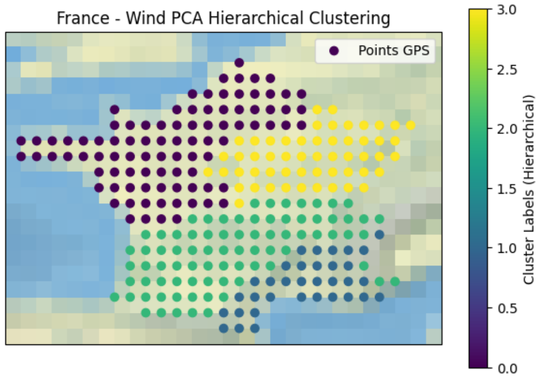
C) Clustering

In this case, we compute and study a segmentation in 4 groups of the wind in France, based on the PCA representation keeping only 10 principal components using Kmeans and hierarchical clustering. Then, we compare the both study (raw data and PCA data) by evaluating Inertia or Sum of Squared Distances (ISS) for the Kmeans clustering and Davies-Bouldin Index for the hierarchical clustering :
- Raw data :
ISS : 8183045.360801062 / Davies-Bouldin Index: 1.6790461671012644
- PCA data : ISS : 846061.4713272876 / Davies-Bouldin Index: 1.6237458985696915
ISS measures how far the points within a cluster are from the centroid of that cluster and the Davies-Bouldin index measures the average similarity ratio of each cluster with the cluster that is most similar to it. A lower ISS for the Kmeans clustering and a lower Davies-Bouldin index for the hierarchical clustering is generally better. We can see a better ISS for the raw data but there isn't a big difference and a better Davies-Bouldin index for the PCA representation. We conclude that PCA representation performs well.







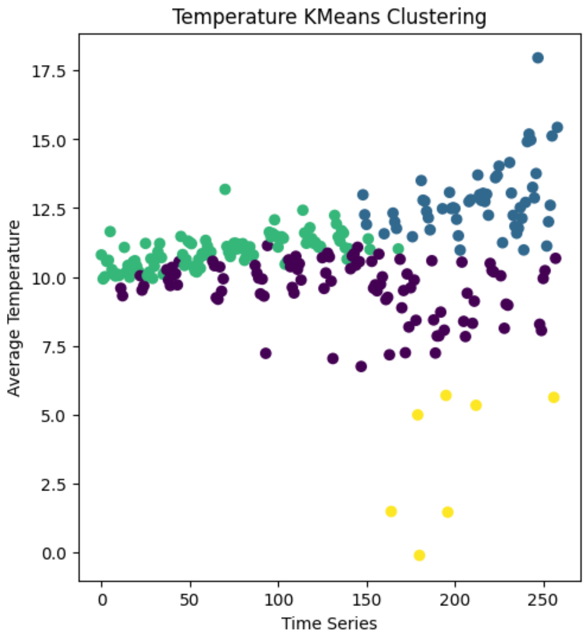

France - Wind PCA Hierarchical Clustering
Points GPS
Cluster Labels (Hierarchical)
3.0
2.5
2.0
1.5
1.0
0.5
0.0

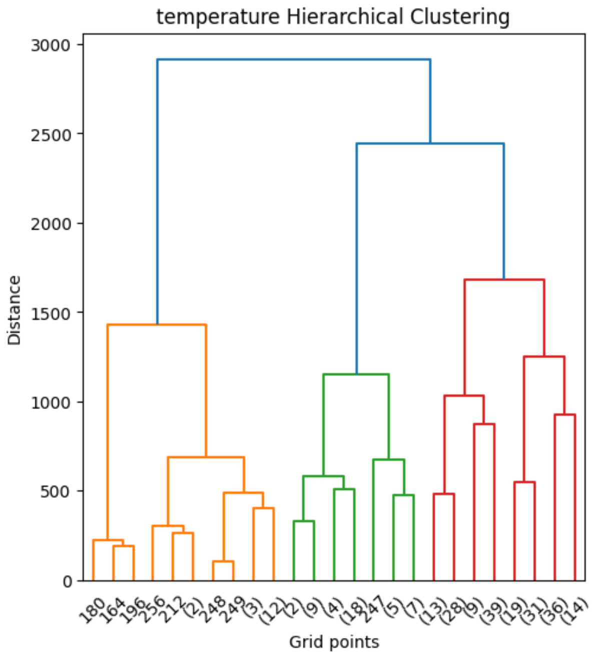
We can see the data points are grouped well together in distinct clusters.

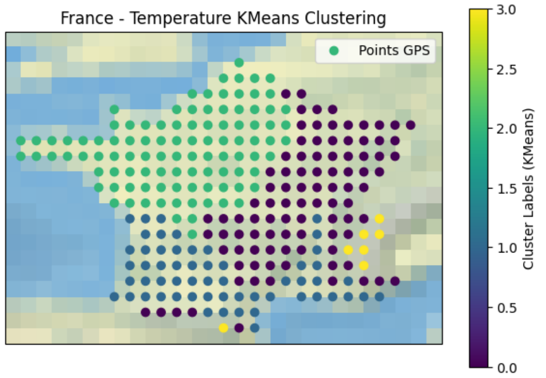
III) Temperature clustering

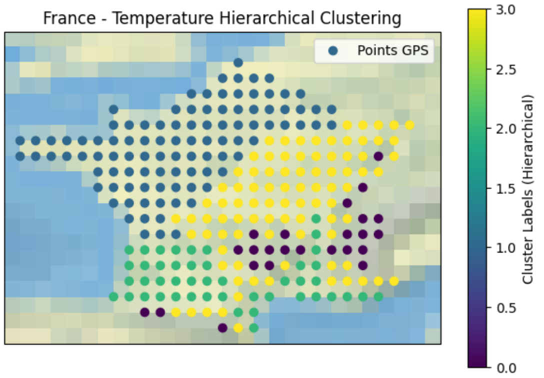
A) Raw data

We do the same study for the Temperature as the wind study. We study and compare the use of the Kmeans and the hierarchical clustering to provide a segmentation into 4 groups of the Temperature using the raw time series.








France - Temperature Hierarchical Clustering
Points GPS
Cluster Labels (Hierarchical)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

We can see a clear separation between the Kmeans clusters and separations between the vertical lines for the hierarchical clustering. We have also the the scores for the Kmeans clustering and the Hierarchical clustering :

ISS : 8093997.268125119 / Davies-Bouldin Index : 3.085335493824959

We have a better ISS but a less good Davies-Bouldin Index than the Wind study.

B) Feature extraction

We do the same algorithm for the Temperature study as the Wind study. The output indicates that we need only 22 components to achieve this threshold. That means 95% of the data are explained by 22 components. We need to keep 22 components because this value is the balance between compression and information retention.

C) Clustering

To perform a segmentation of the temperature time series based on the PCA representation using a model-based clustering method, we can use the Gaussian Mixture Model (GMM) from the scikit-learn library. GMM is a probabilistic model that assumes that the data is generated from a mixture of several Gaussian distributions. Then, we experiment with 4 different covariance types that represent various assumptions about the covariance structure of the Gaussian distributions. For each clustering instance, we need to choose how many clusters we need and we start with a moderate number of clusters like 3 clusters and assess the quality of the clustering.

- Our own and personal choice which is the full model (The default model) :

The full model performs well when there are 5 clusters.

- The spherical model with equal volume :

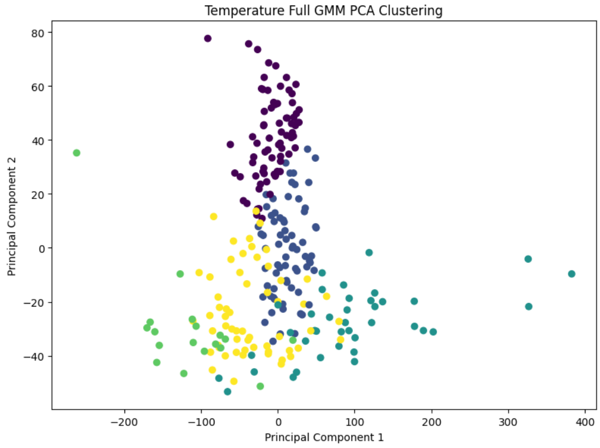
We have a good separation between the clusters when there are 3 clusters for the spherical model with equal volume.

- The spherical model with unequal volume :

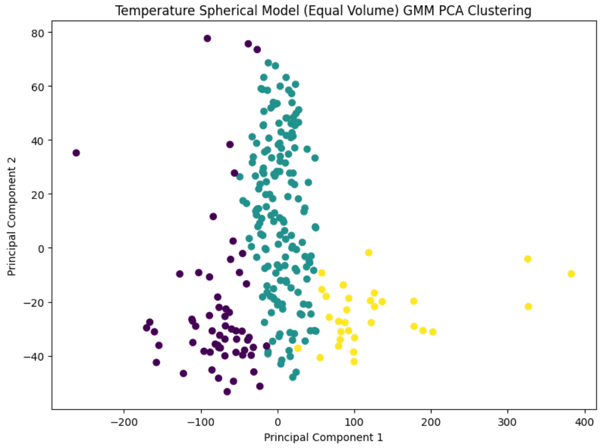
We have a good separation between the clusters when there are 4 clusters for the spherical model with equal volume.

- The diagonal, varying volume and shape :

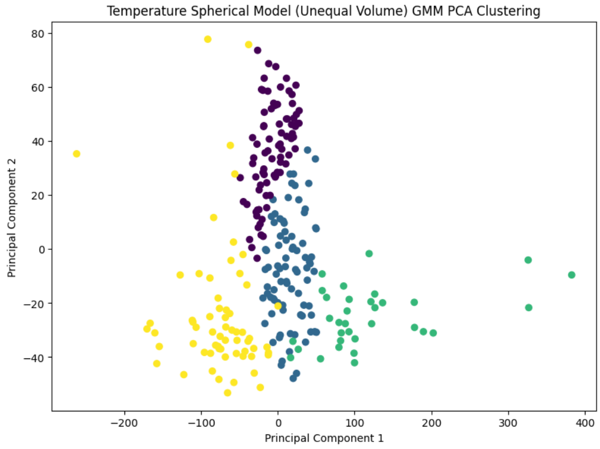
In this case, the clusters can have different shapes, orientations, and volumes. The best number of clusters 5 because there is a clear separation of the 5 clusters. This covariance type allows for more flexibility in cluster shapes.



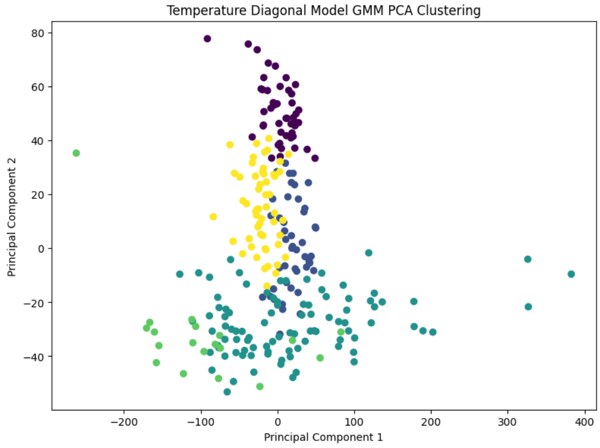
Temperature Full GMM PCA Clustering



Temperature Spherical Model (Equal Volume) GMM PCA Clustering



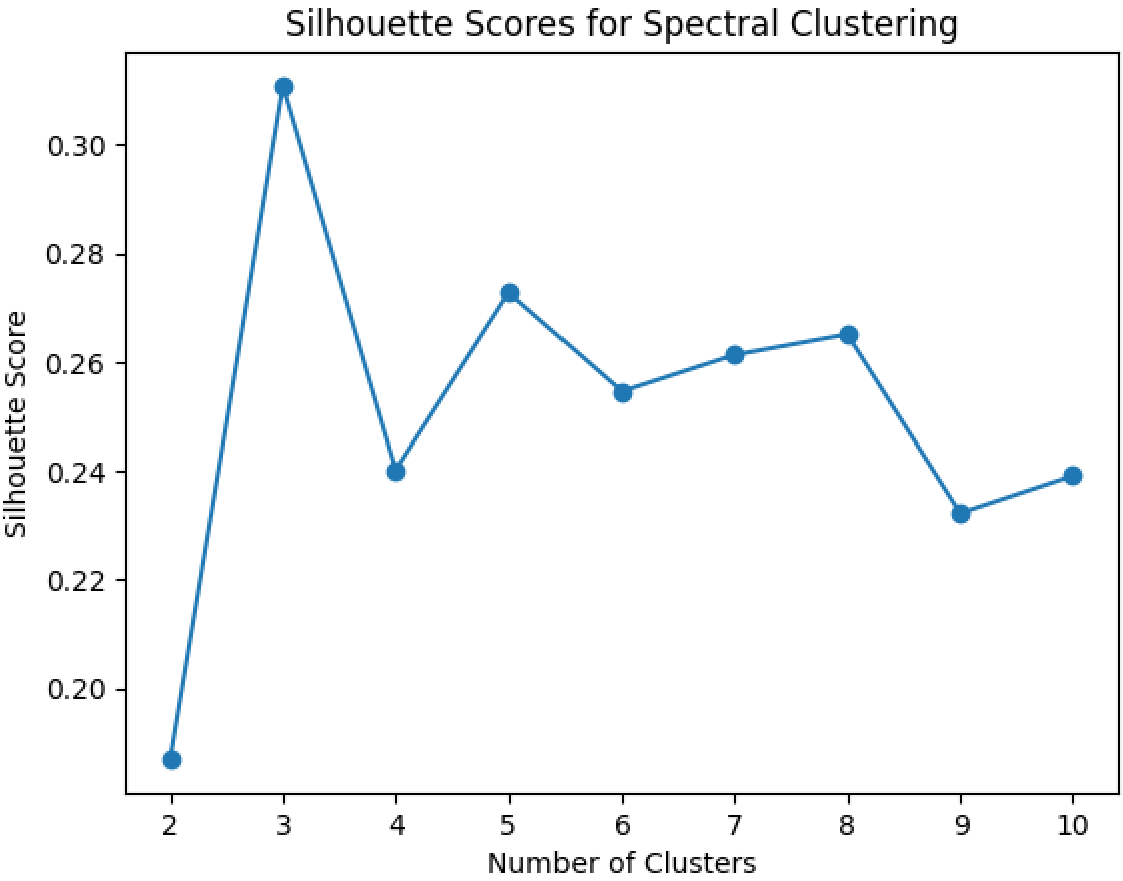
Temperature Spherical Model (Unequal Volume) GMM PCA Clustering

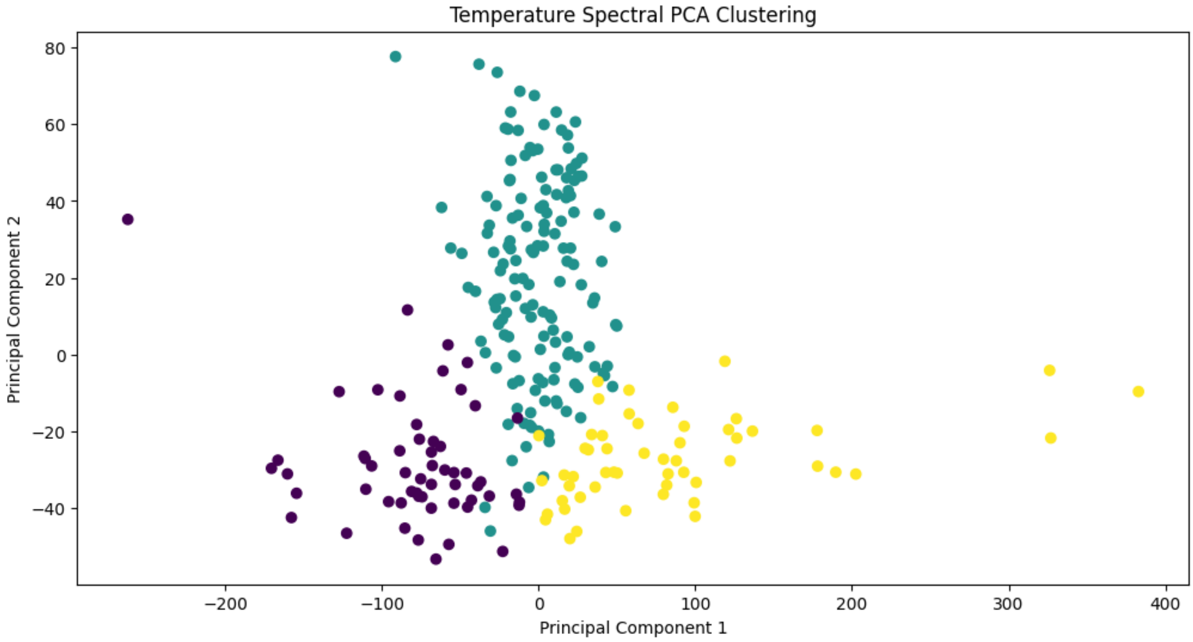


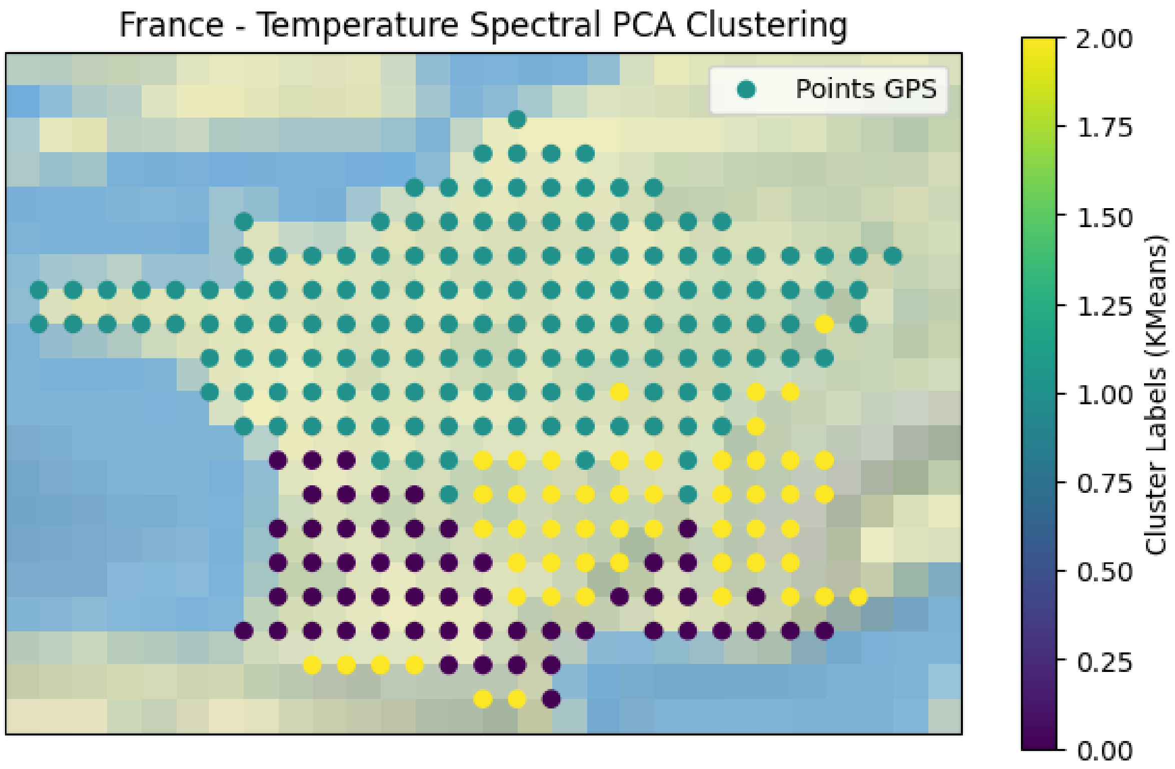
Temperature Diagonal Model GMM PCA Clustering

D) Clustering using spectral clustering

We need to choose the appropriate number of clusters for spectral clustering. A method is to calculate the silhouette score for different numbers of clusters. The silhouette score measures how well-defined the clusters are, and a higher score indicates better-defined clusters. In our case the best number of clusters is 3 clusters.

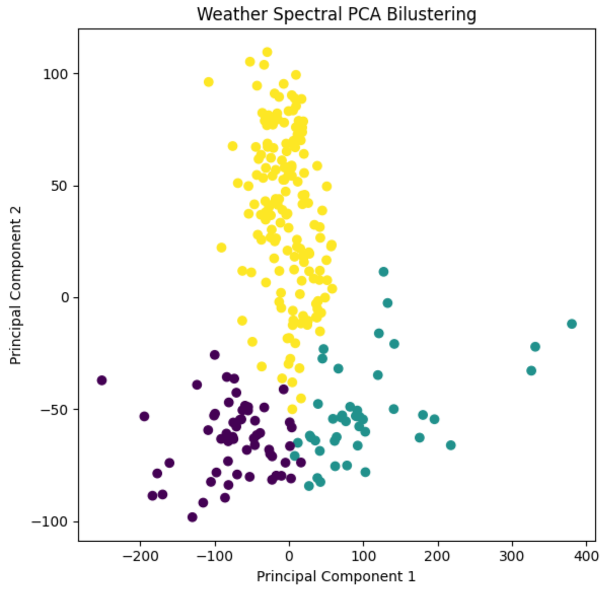

Silhouette Scores for Spectral Clustering
Silhouette Score
0.30
0.28
0.26
0.24
0.22
0.20
2
3
4
5
6
7
8
9
10
Number of Clusters

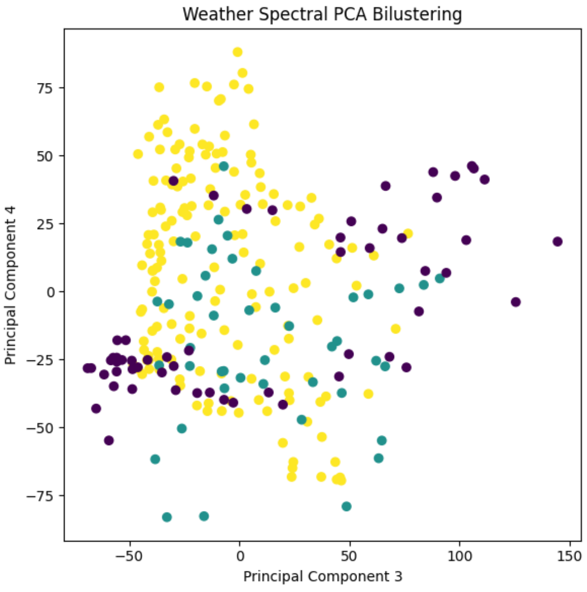

Temperature Spectral PCA Clustering
Principal Component 2
80
60
40
20
0
−20
−40
−200
−100
0
100
200
300
400
Principal Component 1



IV) Temperature and Wind Clustering

In order to propose an approach to cluster wind and temperature data at the same time, we have to concatenate wind and temperature data to create a new dataset with both variables. After that, we standardize the combined data to ensure that both wind and temperature variables are on the same scale and we apply PCA to reduce the number of features. Then, we tried applying Spectral Biclustering but I didn't have any good results.


Weather Spectral PCA Bilustering
Principal Component 2
Principal Component 1


Weather Spectral PCA Bilustering
Principal Component 4
Principal Component 3