Machine learning for clustering

Omar Krichen & Zakaria Benbouzid

January 8^{th}, 2024

1 Introduction

This report explores the intricate world of weather patterns in France, focusing on temperature and wind data from 259 different locations. Weather data can be tricky to understand because it's vast and complicated. We want to overcome these challenges by using advanced methods that can uncover hidden patterns in the information.

Traditional ways of looking at raw weather data often fall short, so we're exploring new and more efficient methods. Our goal is to find meaningful structures in the data, which will help us understand weather patterns in a more detailed way.

To start, we've picked some key cities in France to ground our analysis in real locations. The report then dives into different ways of looking at temperature and wind data, addressing issues like too much data, understanding the information, and dealing with different assumptions. We compare these methods to see which ones work best.

By doing this, we aim not only to improve our understanding of weather patterns but also to contribute to how researchers, weather experts, and decision-makers can benefit from these advanced methods for a deeper insight into regional weather.

2 Explanation of the data

The data at the heart of this analysis comprises time series information on temperature and wind patterns collected from 259 distinct grid points across the French territory. Each grid point represents a unique location, forming a comprehensive network that spans the entire geographical landscape of France. For each grid point, hourly measurements were recorded over the course of a year, resulting in a rich dataset of 8760 hours that encapsulates the temporal evolution of weather conditions.

To complement the temporal aspect of the data, the geographical coordinates (longitude and latitude) of each grid point are included. This spatial information is instrumental in contextualizing the weather patterns, allowing for the correlation of climatic variations with specific regions. By leveraging this comprehensive dataset, the analysis aims to uncover underlying patterns, relationships, and clusters within the weather data, offering valuable insights into the dynamics of French weather.

3 Methods to analyze the data

KMeans and Hierarchical clustering are two widely used techniques in data analysis, each offering distinct advantages and presenting unique challenges when applied to weather data, specifically wind and temperature time series.

KMeans is a partitioning-based method that groups data into a pre-defined number of clusters. One of its primary advantages is computational efficiency, making it suitable for large datasets like hourly temperature and wind measurements from numerous grid points. KMeans is straightforward to implement and can provide clear, non-overlapping clusters, which can be advantageous for straightforward interpretation. However, a notable inconvenience lies in its sensitivity to initial cluster centroids, and the need for specifying the number of clusters beforehand can be challenging when the optimal cluster count is uncertain. Additionally, KMeans assumes clusters of similar size and shape, which may not always align with the inherent variability in weather patterns.

Hierarchical clustering, on the other hand, builds a tree-like hierarchy of clusters. One notable advantage is the flexibility it offers in exploring different levels of granularity, allowing for the identification of both broad and fine-scale patterns in the weather data. Hierarchical clustering does not require pre-specifying the number of clusters, addressing a significant limitation of KMeans. However, this flexibility comes at a cost in terms of computational complexity, especially for large datasets. The hierarchical structure may also make interpretation challenging, and the method is sensitive to noise, potentially leading to the formation of spurious clusters.

Comparing the two methods, KMeans may be more suitable for quick exploratory analyses, providing a computationally efficient way to identify clear clusters. However, its reliance on a pre-defined number of clusters and sensitivity to initial conditions may limit its effectiveness. Hierarchical clustering, while computationally more demanding, offers a more nuanced exploration of the data, accommodating varied cluster sizes and shapes. It is particularly useful when the inherent structure of the data is not well-known, allowing for a more flexible and adaptive approach to cluster identification.

Principal Component Analysis (PCA) is a powerful technique employed to reduce the dimensionality of complex datasets, and its application to wind and temperature time series data from 259 grid points in this analysis is no exception. The high-dimensional nature of the time series, with hourly measurements recorded over a year for each grid point, can pose challenges in terms of interpretability and computational efficiency. PCA mitigates these challenges by transforming the original variables into a new set of uncorrelated variables called principal components.

The advantages of applying PCA to wind and temperature data are multifold. Firstly, it enhances interpretability by expressing the data in terms of a smaller set of components that retain the essential information. This reduction facilitates a clearer understanding of the dominant factors influencing the weather patterns. Secondly, PCA aids in computational efficiency, especially when dealing with large datasets, as it focuses on the most impactful features,

streamlining subsequent analyses. However, it's crucial to note that PCA assumes linearity and may not capture non-linear relationships effectively.

4 Results of the analysis

To discover our data, we chose to examples for 3 different cities in France : Brest,Lille and Lyon.

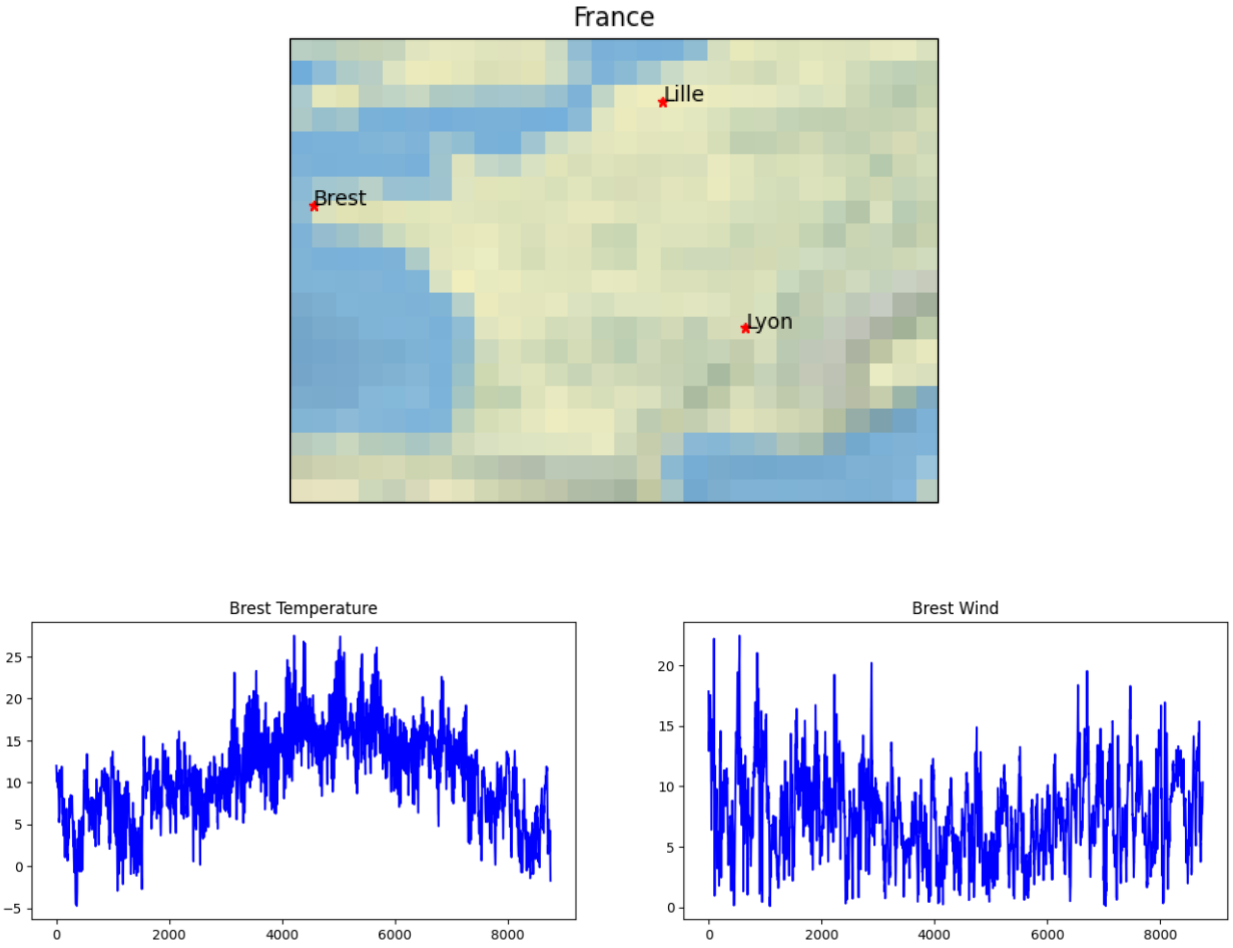

Figure 1: Weather in Brest

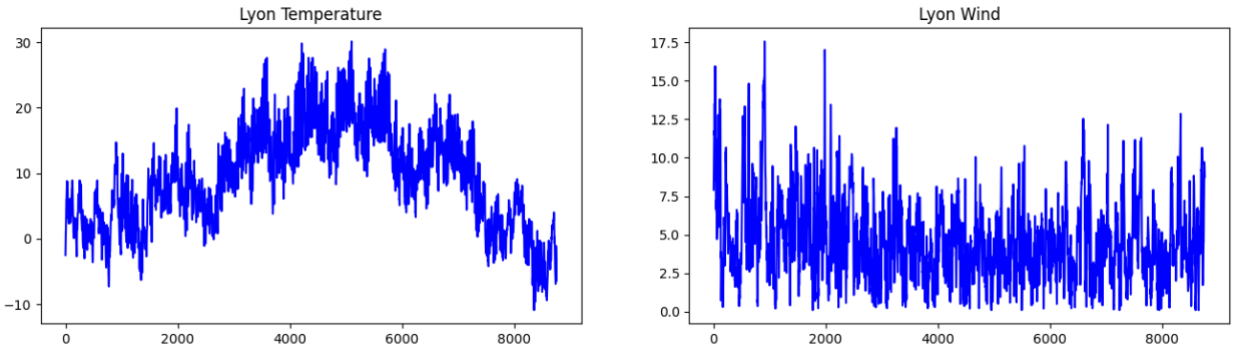

Figure 2: Weather in Lyon

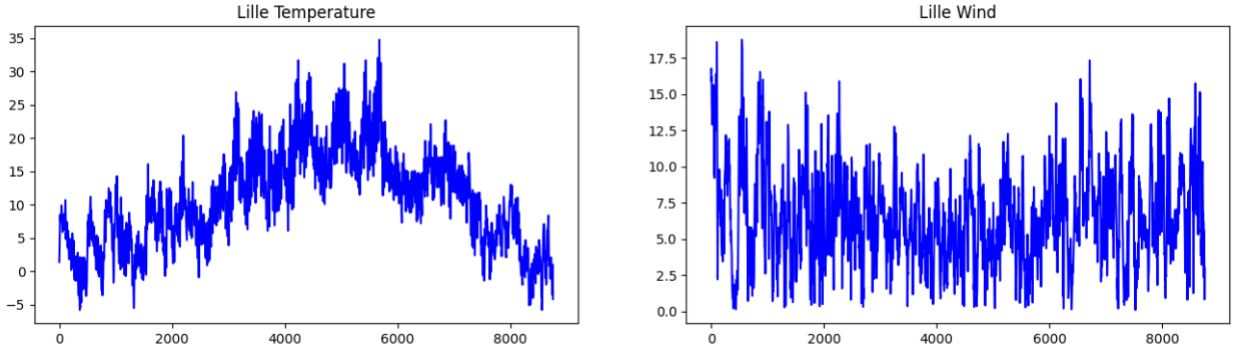

Figure 3: Weather in Lille

A Naive method for segmentation of the weather is to use the average of the Timeseries of Wind and Temperature for the whole period.We tried to cluster it using KMeans and Hierarchical Clustering.

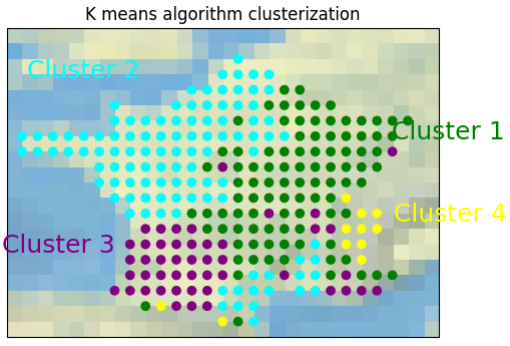

Figure 4: KMeans Clustering

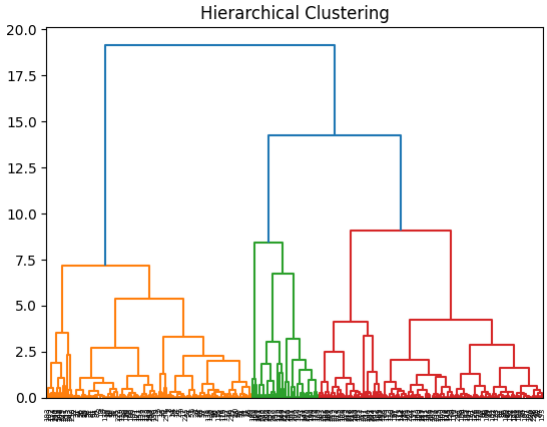

Figure 5: Hierarchical Clustering

Unfortunalety, this results aren't very precise for decision makers,for that we will search for others methods.

Now,we will work on the Wind data . We applied the clustering algorithms with a purpose to compare their efficiencies using the **Silhouette Score** with 4 clusters:

-KMeans Silhouette Score: 0.197

-Hierarchical Silhouette Score: 0.167

The silhouette score measures how well-defined the clusters are. A higher silhouette score indicates better-defined clusters.For that,we concluded that apparently KMeans algorithm is better in our case. (We had the same conclusion in Temperature data)

To reduce the complexity of our Timeseries, we will try to reduce dimensions using the **PCA** technique. Our aim here is to reduce the n=259 cities.

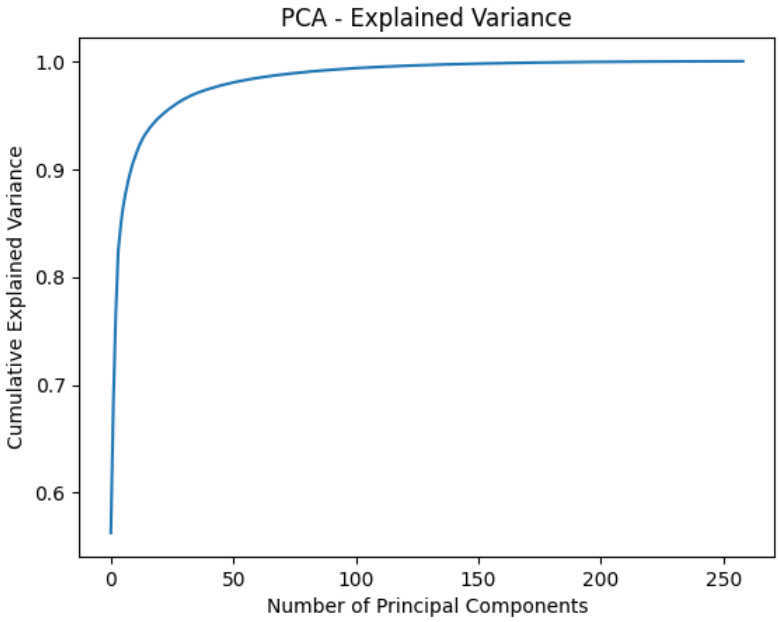

Figure 6: PCA - Explained Variance

For 95% variance achieved,we need to keep 28 principal components in the wind data comparing to only to 2 principal components in the temperature data. We can see that complexity can be reduced considerably with the PCA technique because 95% precision is already perfect for our analyses comparing to the cost of treating 259 cities.

To approve that **PCA** technique is efficient, we will try to study our data with reducing its dimension to n=10.

Comparing the silhouette score of KMeans and Hierarchical clustering, we had for wind data:

-KMeans Silhouette Score: 0.418

-Hierarchical Silhouette Score: 0.402

We can see that kMeans is always having better result, but the important thing that we are having better scores comparing to n=259.

For Temperature, we used a model based clustering method called **Gaussian Mixture**.

We tried to compare different densities (spherical model with equal volume , spherical model with unequal volume, diagonal, full) using this time the **BIC** (Bayesian information criterion) aiming to find the best number of cluster for each density.

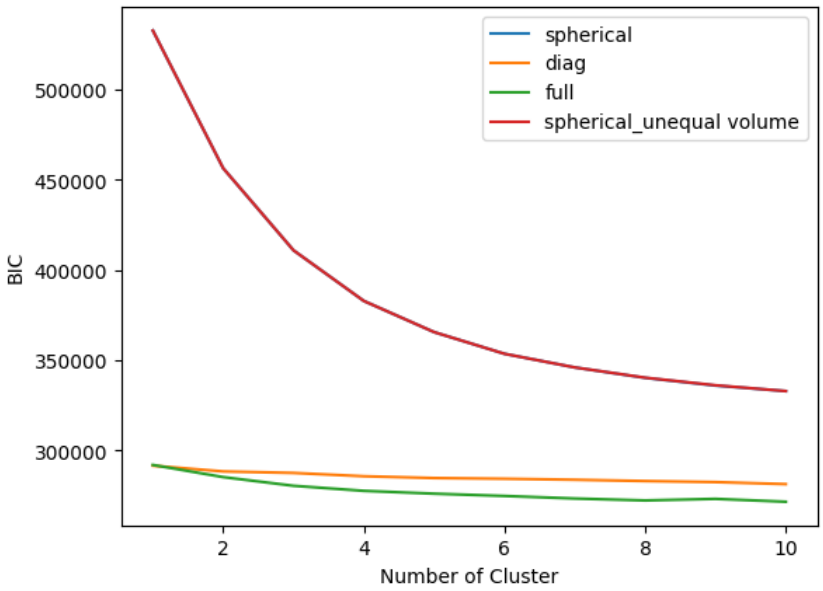

Figure 7: Gaussian Mixture

We suggest to use 4 clusters for Spherical clustering because the BIC doesn't vary significantly anymore after, and 2 clusters to diag and Full for same reason. Like that he are having to least clusters with better efficiency.

We tried also another method called **Spectral Method** aiming to find the best number of cluster for temperature data.

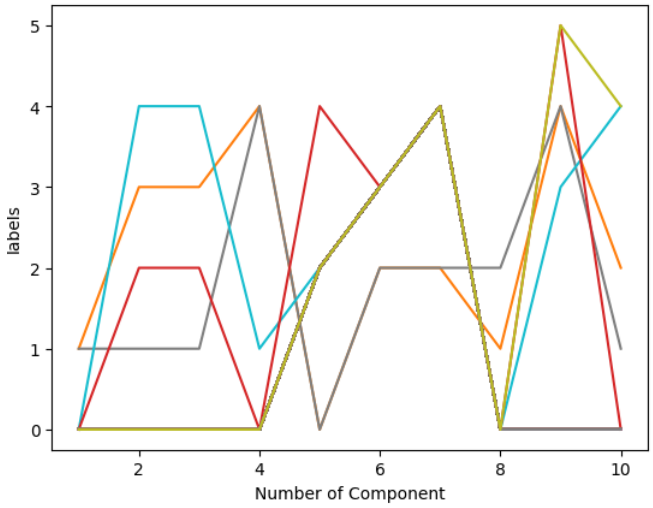

Figure 8: Spectral Clustering

Unfortunately, it is really hard to conclude which number to take.

The goal right now is to work on clustering temperature and wind in the same time. For that we need to work like in 3 Dimensions.
So, we propose to regroup the Timeseries in 1 Dataframe with the function **np.dstack** and then we use the **TimeSeriesKMeans** to fit our model with KMeans since it's the algorithm that worked well with our data.

5 Conclusion

In summary, the choice between KMeans and Hierarchical clustering depends on the specific objectives of the analysis. KMeans may be preferable for quick, computationally efficient insights, while Hierarchical clustering provides a more flexible, granular exploration of the data, albeit at a higher computational cost. Both methods contribute valuable perspectives to the understanding of weather patterns, and the choice should align with the desired depth and interpretability of the analysis.