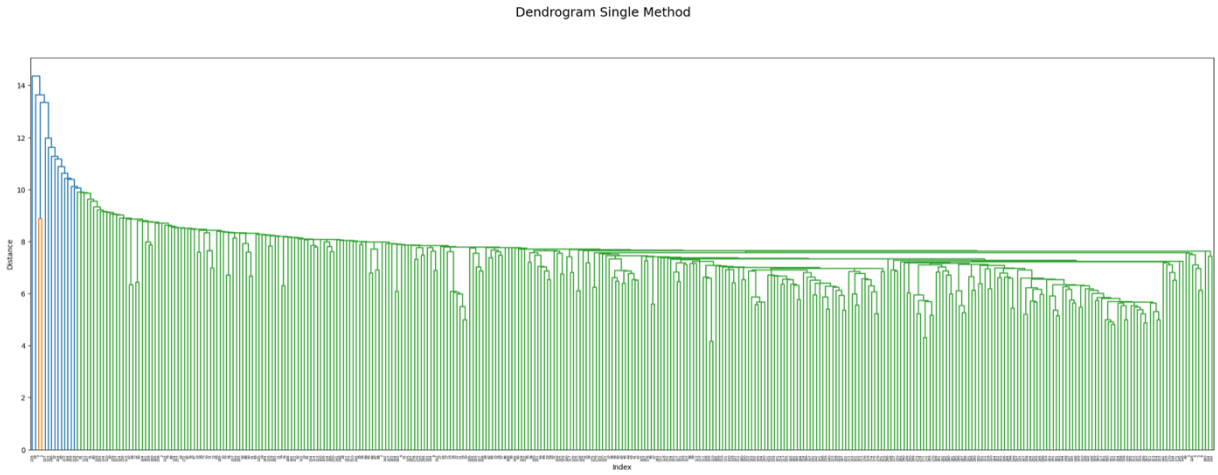
Giada Griso
September 1, 2024
Machine Learning Course
Exercise 2.1

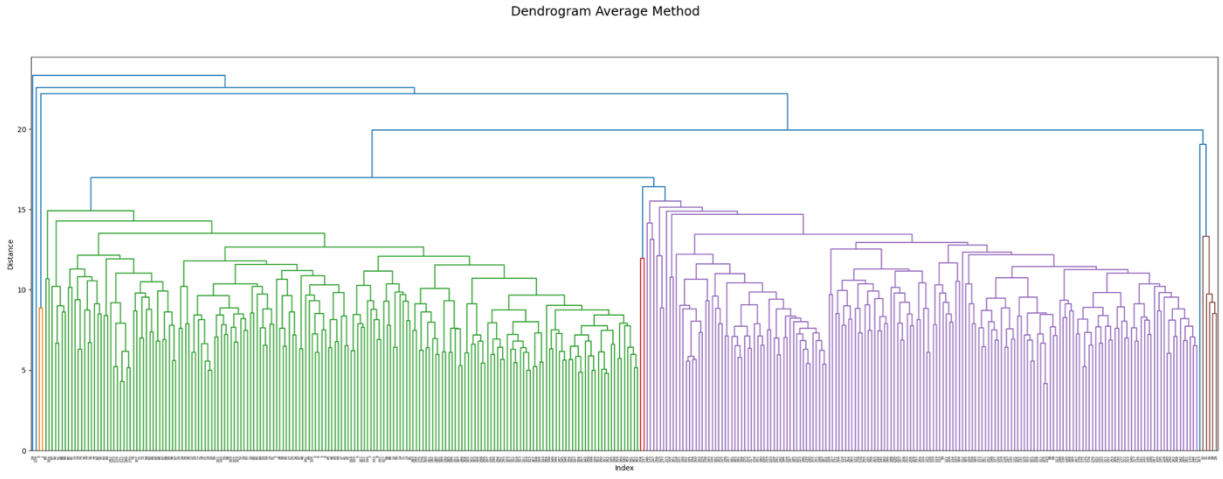
Unsupervised Learning Algorithms

1. Dendrograms

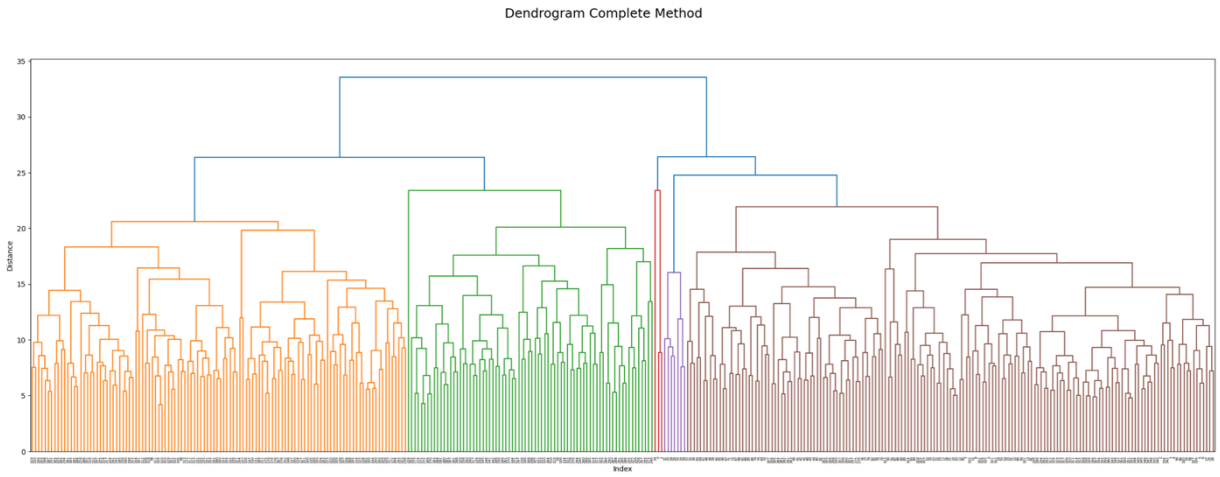
The four different dendrograms below show how the weather data from 15 different European stations in **2015** can be grouped together. Each dendrogram uses a different method to figure out how these stations might be similar to each other.



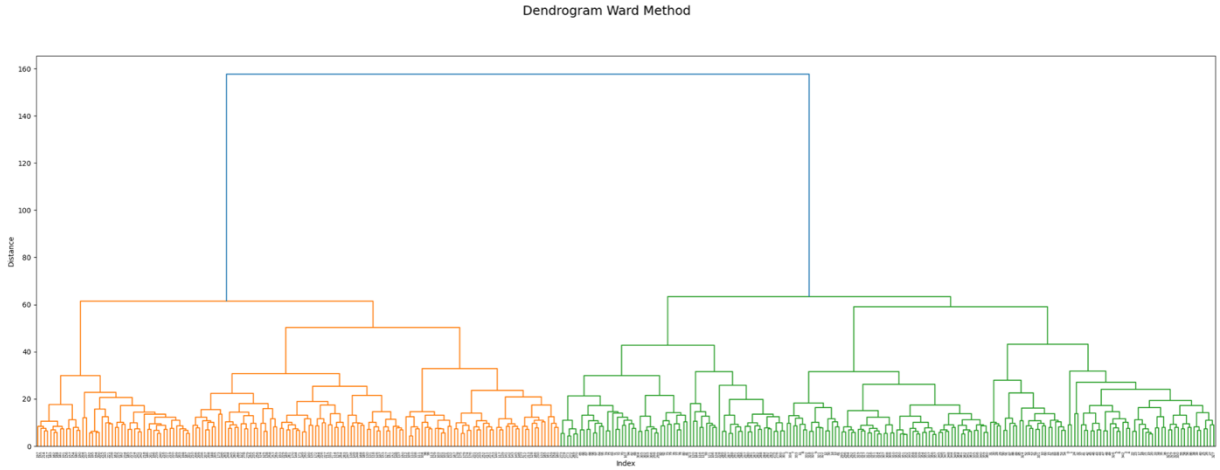
The **single method** generates a very dense dendrogram, with many small clusters merging at very low distances, indicating less distinct separation among the data points. The method appears to be noisy and chaotic with weather data due to its sensitivity to small variations.



The **average method** averages the distance between clusters and tends to balance between the minimum and maximum distances. The dendrogram displays more structure with clearer groupings (4-6), but the clusters are still somewhat intermingled, suggesting that some variability hasn't been effectively separated.



The **complete method** considers the maximum distance between points in different clusters, leading to compact clusters. It shows larger, more compact clusters but still has + multiple groups merging at higher distances, indicating some clarity but potential overlap among clusters. Here we can see approximately 3-5 clusters, as large sections are grouped together.



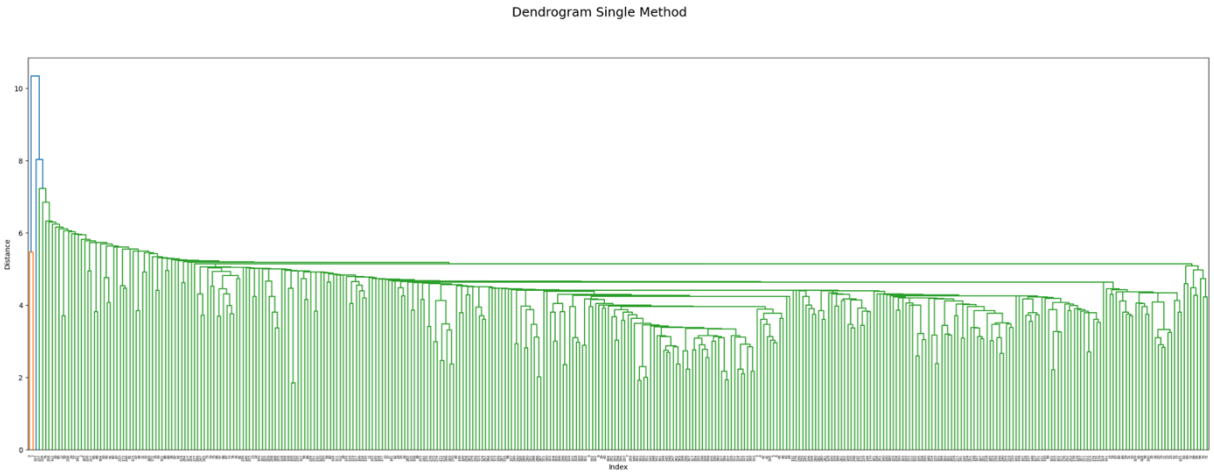
The **ward method** minimizes the variance within each cluster, trying to keep the groups as neat and similar as possible. Among all methods, it is the one showing the most distinct and obvious groupings, with only 2 main big clusters likely representing broad climate zones across Europe (e.g.: northern vs. southern regions).

Practical implications

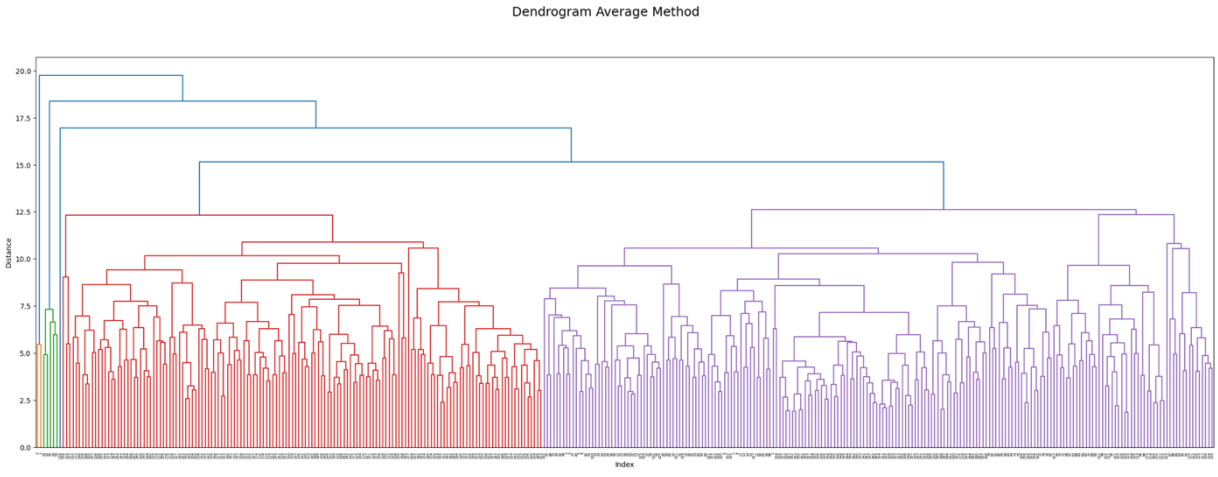
The single method is more suitable to identifying many small, nuanced weather patterns, though it can be overwhelming. The average and complete linkage methods offer a more balanced view, showing clusters of stations that share moderately similar weather characteristics without too much complexity. In contrast, the ward method provides the clearest big-picture view by identifying the most distinct and obvious groupings, making it ideal for understanding broad regional weather patterns across Europe.

2. PCA Reduced Datasets with 11 Components

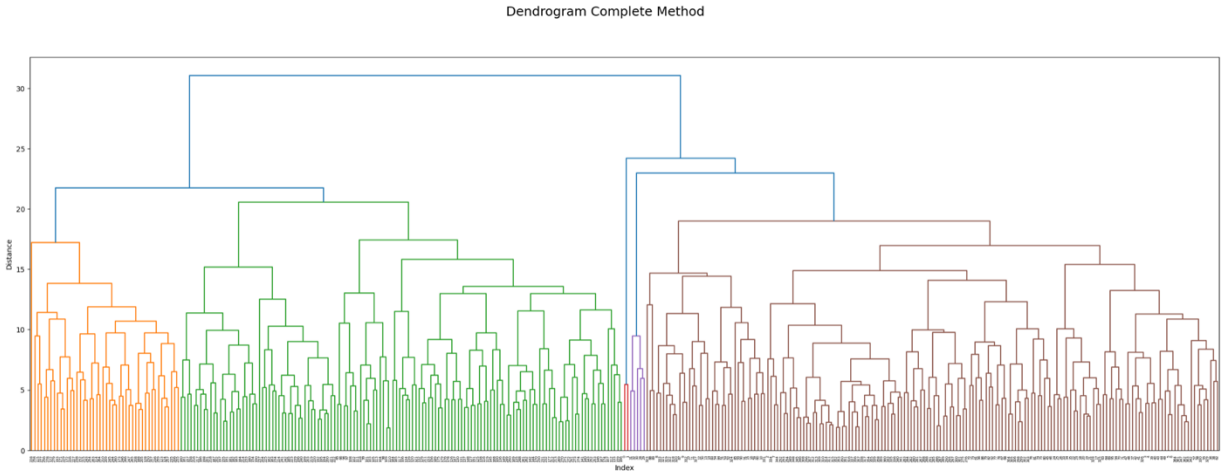
These are the dendrograms we obtained after performing a PCA and reducing the analysis components to 11 only, each of them representing one of the recorded weather parameters:



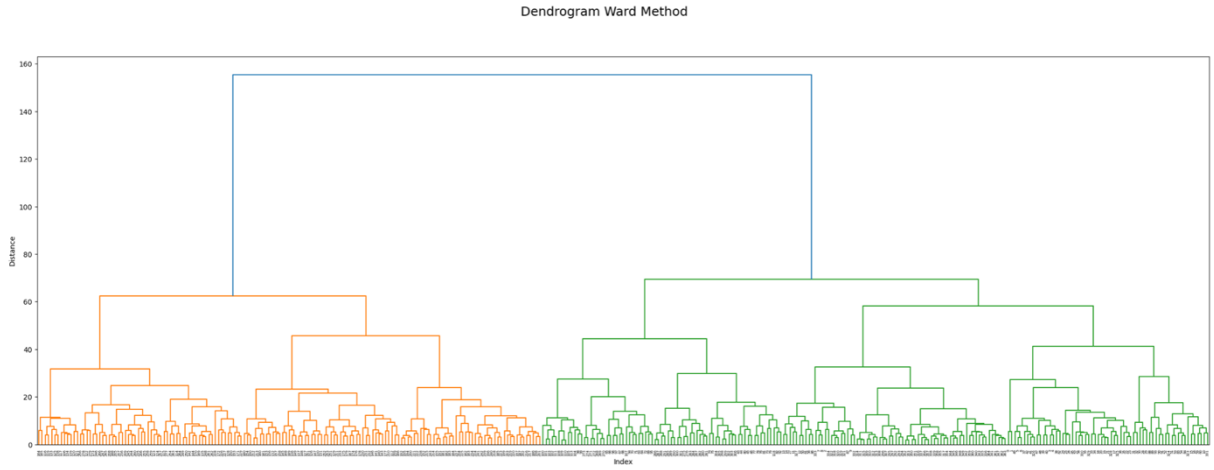
The dendrogram appears a bit more structured compared to before. The reduction in dimensionality has likely removed some noise, allowing for more distinct clusters to form. Single linkage still tends to produce elongated clusters though.



The dendrogram after PCA shows a clearer separation of clusters, with more distinct groupings that are better spaced apart. This suggests that PCA has helped in isolating the most significant patterns, making the clusters more meaningful.



There's a noticeable improvement in the clarity of the clusters. The dendrogram now shows more distinct and separated clusters with reduced overlap, indicating that PCA has effectively condensed the data to its most relevant features, allowing for clearer cluster formation.



The ward method doesn't show any significant improvement after performing the PCA.

Conclusions

The clustering after PCA is definitely more conclusive, especially for the complete method. The dendrograms show clearer, more distinct clusters with less noise and overlap, indicating that the PCA has successfully highlighted the key patterns in the data.

To see if any of these clusters match "pleasant weather" labels, we would need to overlay the labels on the clusters to see if one or more of them predominantly contain pleasant weather data.

The main 2 clusters obtained by the ward method could also represent different seasons, considering that the weather patterns in Europe have strong seasonal components. For example, one large cluster might correspond to winter months, while another could correspond to summer months.