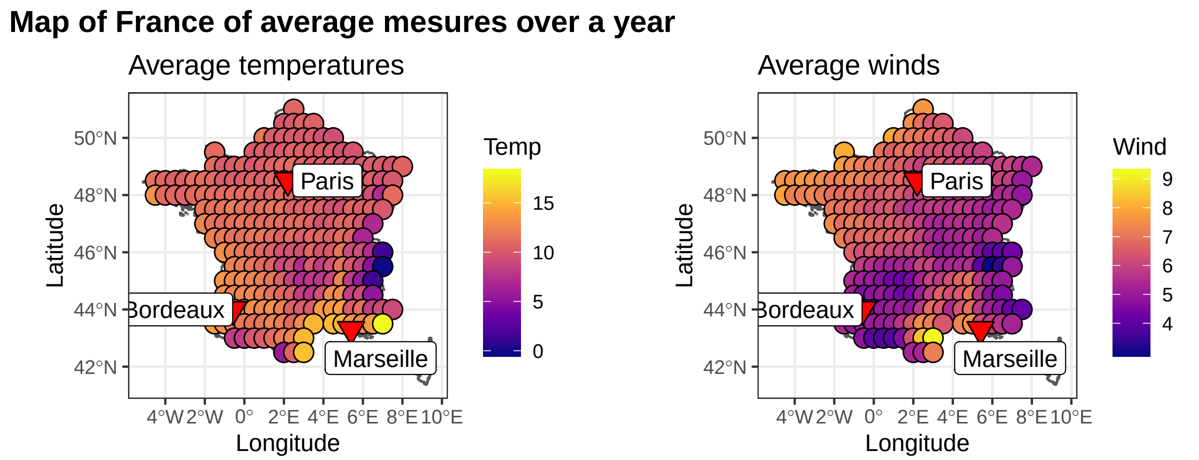
MAL 2 Project

Frederick Deny, Paul Canat

10/12/2019

Introduction



Map of France of average mesures over a year

Average temperatures

Average winds

In this project we will study the segmentation of a Temperature and Winds dataset, in the effort of segmenting France into climate groups. The dataset contains 259 observations and 8760 features, which are in fact measurements taken every hour during a year.

As a clustering quality indicator we will use the silhouette indicator, which evolves between -1 and 1, being closest to one the best clustering.

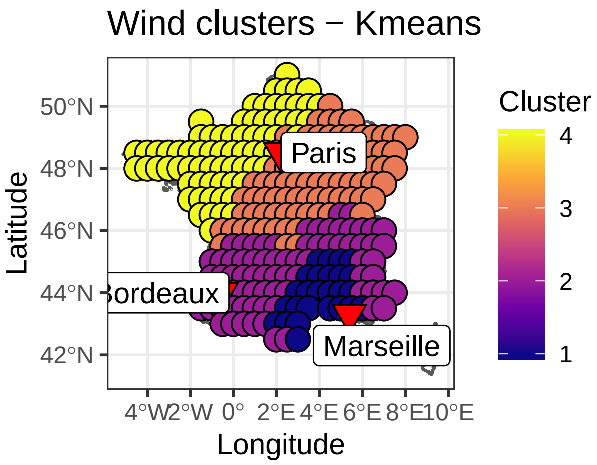
Wind

Raw Data

K-means

Silhouette:

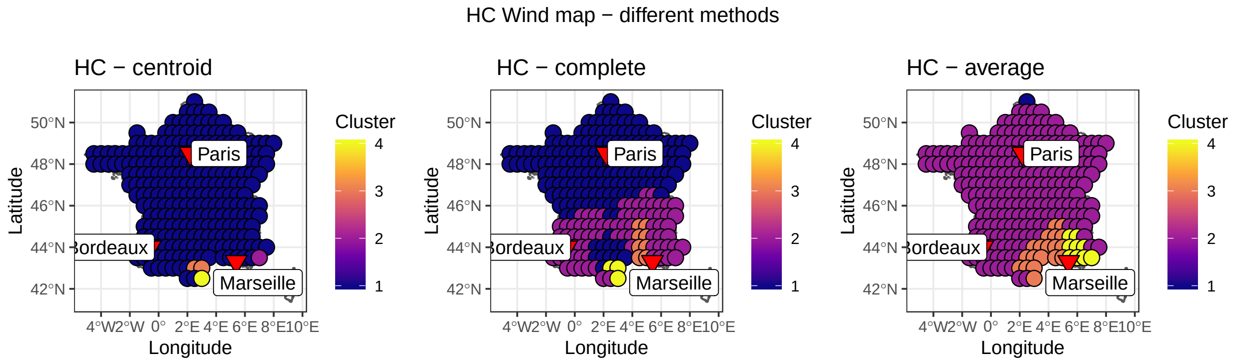
```
## [1] 0.1693398
```

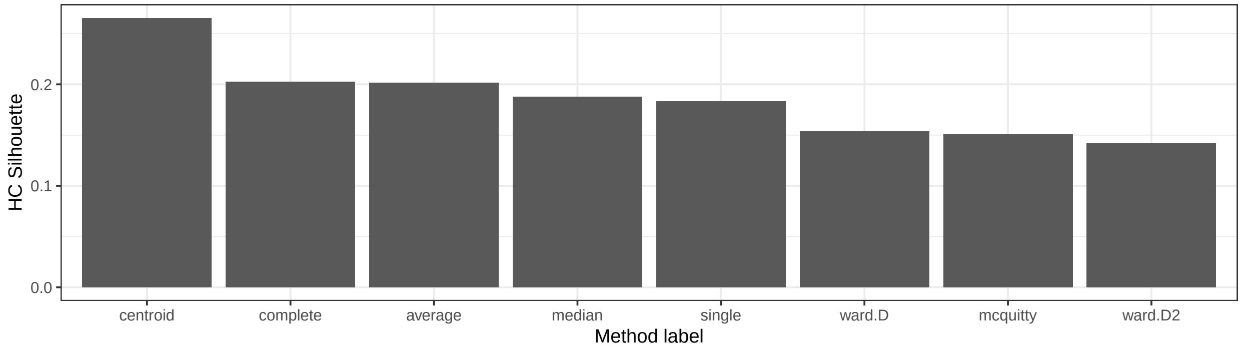


We obtaine here are rather weak silhouette result but the plot shows a make which makes sense considering France's geological caracteristics.

Hierarchical clustering

Distance method examination

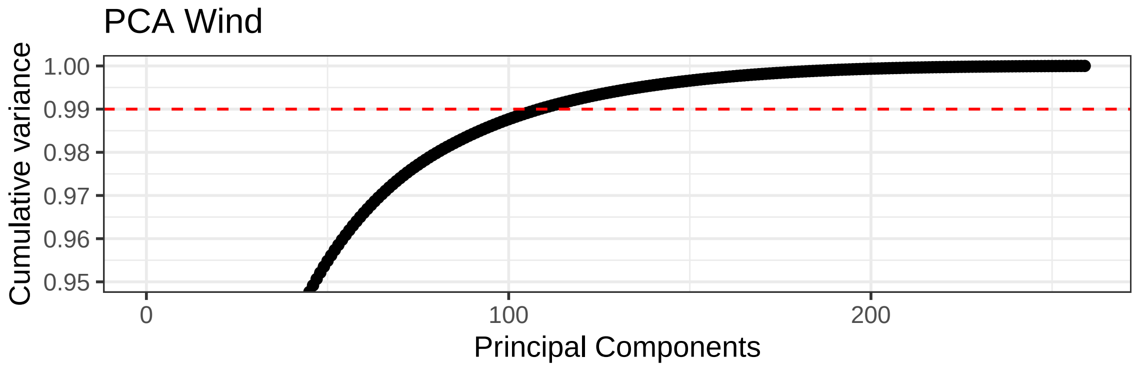




The overall results are better than with the kmeans, but we only plotted the best and we see that our best result gives us a rather unsatisfactory clustering, meaning we will keep the the second best, corresponding to the "complete" method of cluster aggregation.

We initially believed that with more clusters, the "centroid" method would prevail, but after testing the observation distribution remains inadequate.

Feature extraction



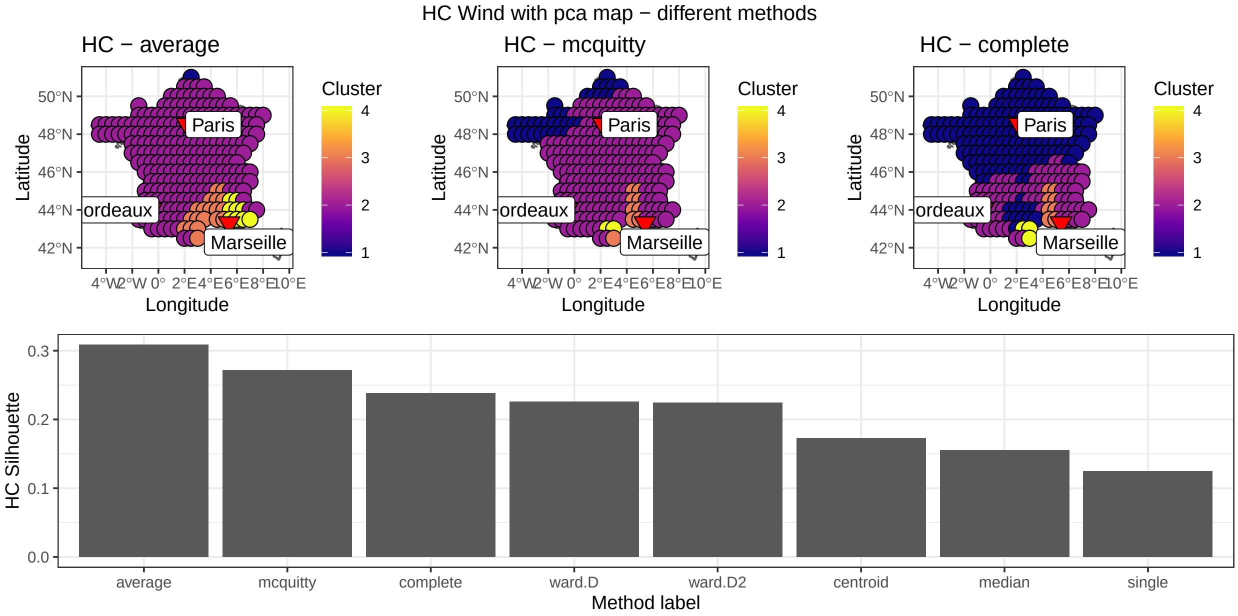
By keeping the best 109 components, we keep more than 99% of our variance, with twice as less observations.

Following the subject, we kept only 10 variables and with the kmeans clustering we got this value:

```
## [1] 0.170787
```

Which is slightly better but remains lower that the hierarchical clustering.

The resulting map is almost identical of the first one, so it won't be shown here.



Here again we only ploted the best results, with better silhouette values overall, the maps are quite less diverse but remain believable.

In the wind clustering, I believe the most convincing result to be the hierarchical clustering with the "average" aggregating method, eventhough the resulting map lacks an equilibrium between the clusters.

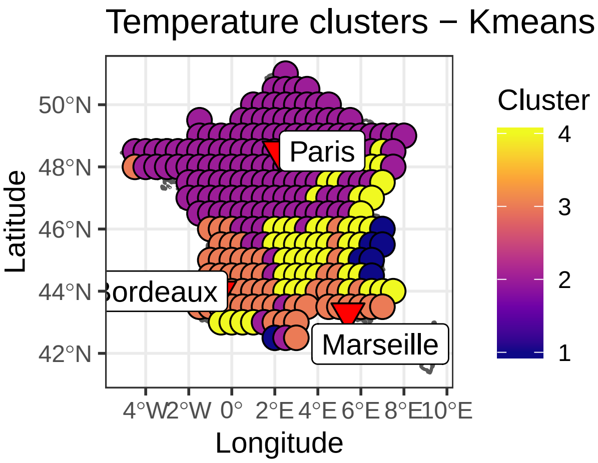
Temp

Raw Data

K-means

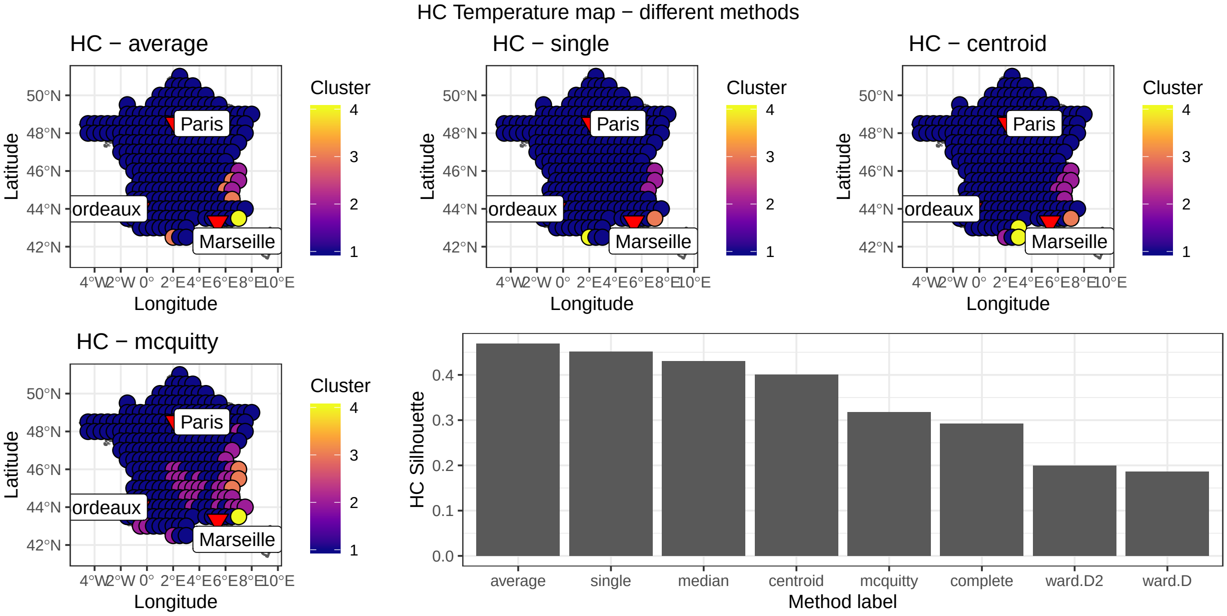
We have an acceptable silhouette value compared to the wind dataset:

```
## [1] 0.2503331
```



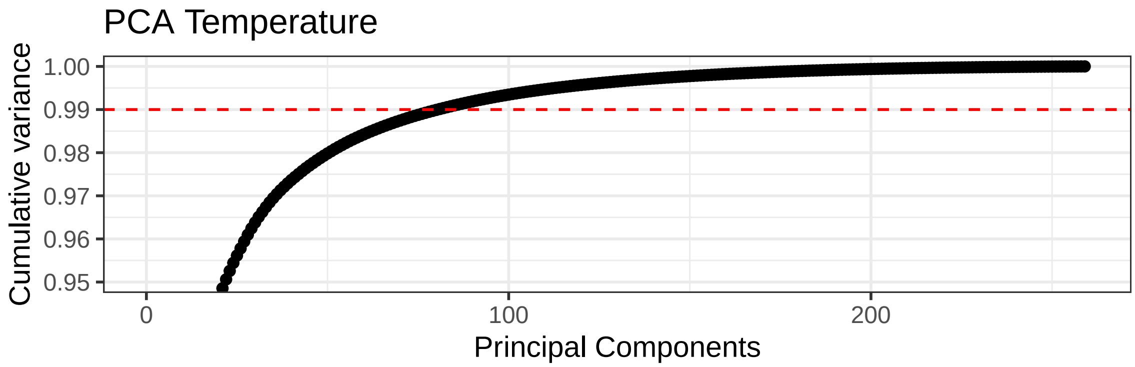
And the maps interestingly enough seems to represent quite accuratly the different geological assets of France: the different mountain ranges being in cluster 3, the Loire and Rhône Basins in cluster 2, some parts of the highest points in cluster 4 and the plains in cluster 1.

Hierarchical clustering



Here the silhouette values are the highest encountered but the maps aren't satisfactory, showing again a lack of diversity. The "mcquitty" might be our best choice here.

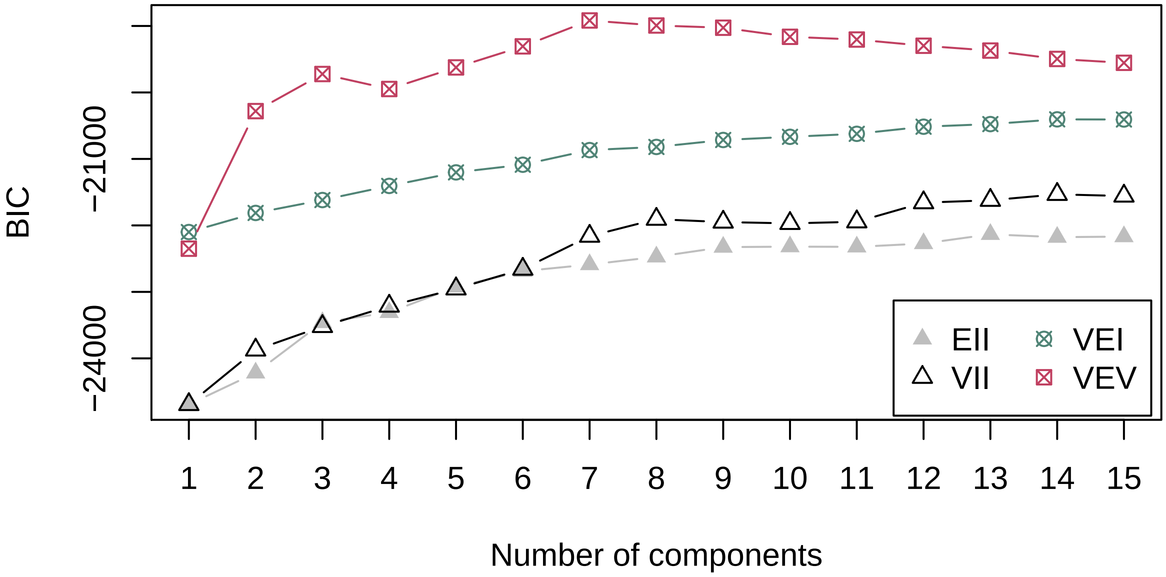
Feature extraction



Again we chose the principal components allowing to retain 99% of our initial standart deviation.

This time we use our pca to considerably reduce our Mclust() function's runtime.

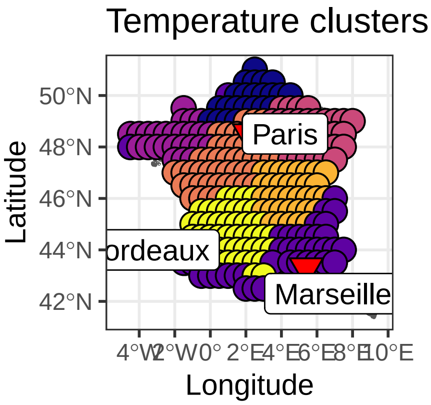
After testing the mclust's different gaussian model we kept the VEV one: ellipsoidal with variable volume and equal shape, while keeping 9 clusters. It was indeed the caracteristics that maximised the BIC.



But even with that maximisation the resulting silhouette value is quite low:

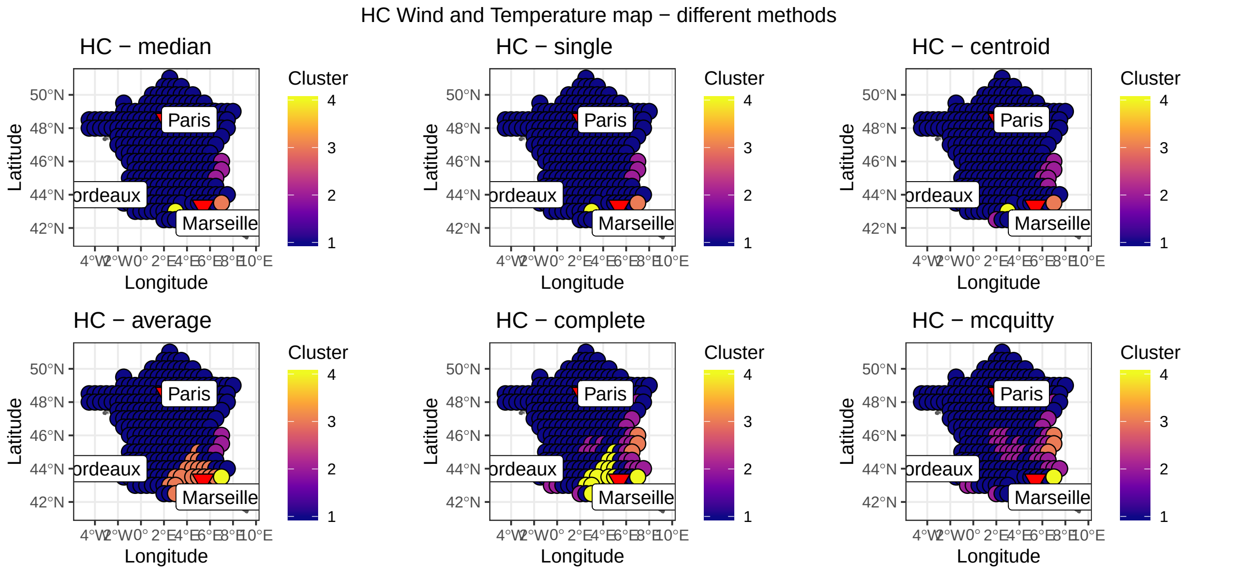
```
## [1] 0.1198353
```

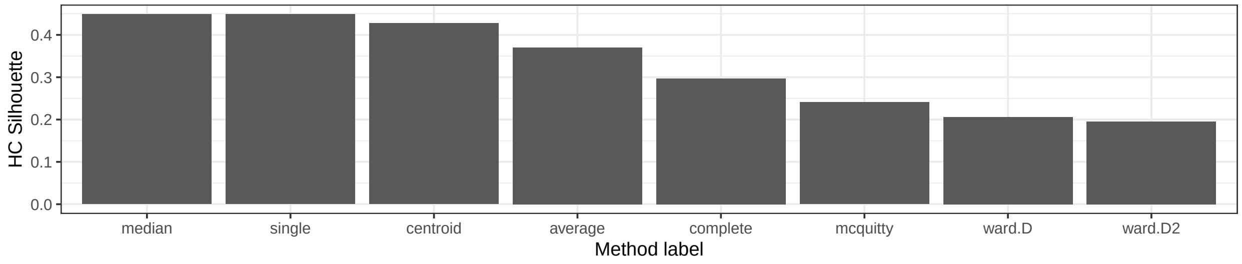
But the map is much more segmented than the ones we produced earlier:



Wind and temperature clustering

We concatenated the datasets, did a pca on them and tested our different HC methods, and obtained these results:





Conclusion

In this project we mainly relied on the silhouette indicator and the layout of the clusters to evaluate a model, but maybe it was a mistake as those criteria sometime went against each other, resulting in confusion and difficulty in selecting the best clustering instances.