Projet portant sur la segmentation du territoire français (métropole)

Mohamed TOUNSI & Mbaimou Auxence NGREMMADJI

20 Decembre 2019

Introduction

Le but de ce TP est segmenter le territoire français en fonction des mesures de temperatures et des mesures de vents. Nous allons utilisé des methodes de clustering afin d'obtenir la meilleure segmentation possible. Un des objectif supplementaire est de faire la segmentation sur le vent et la temperature en même temps.

I Préliminaire

Nous choisissons les 3 villes, en plus de Paris :

	Long	Lat
Paris	2.330	48.860
Lille	3.067	50.633
Nantes	-1.554	47.218
Grenoble	5.725	45.189

(a) Villes choisies

(b) Carte du territoire français

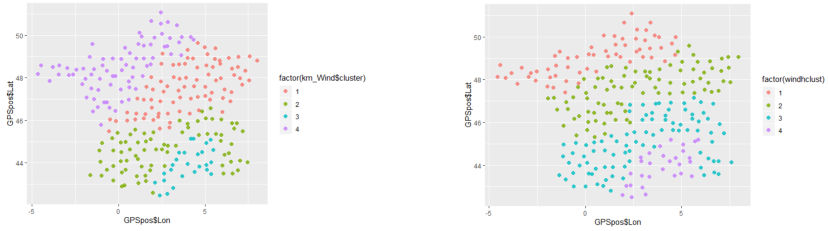
FIGURE 1

II Clustering en fonction des mesures du vent

Dans cette section nous allons procéder au clustering des données concernant le vent. Nous verrons 2 méthodes de clustering, kmeans et le clustering hierarchique.
Nous avons choisi de prendre 4 cluster grâce à la méthode du coude (voir markdown).

II.1 Clustering avec kmeans et hclust sur les données initiales

Nous avons centré les données sans réduire (*scale=F*) car nous avons la même unité de mesure pour chaque colonne (l'ordre de grandeur est le même). Pour la méthode de clustering hierarchique, nous avons utilisé l'argment *method = Ward.D2*. Le choix de la méthode *Ward.D2* s'explique par le fait qu'elle minimise l'inertie intra-classe et maximise l'inertie inter-classe.



(a) Résultat kmeans (b) Résultat hclust

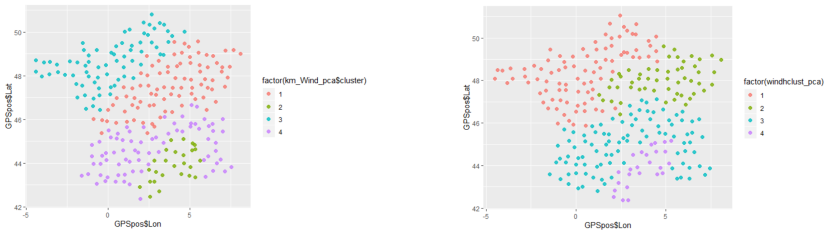
FIGURE 2 – Résultat clustering sur les mesures de vent

Remarque : On constate que les deux méthodes de clustering renvoient quasiment la même segmentation à quelques differences (villes) près surtout du sud-ouest au nord-est.

II.2 Analyse en composantes principales

Nous continuons notre étude en faisant une réduction des dimensions à l'aide de l'analyse des composantes principales (ACP).
Notre choix personnel est 15 composantes qui correspondent à 85% de variances cumulées c'est à dire qu'on conserve 85% de l'information initiale. Avec 10 composantes principales nous avons accès à 81% de variances cumulées. Nous retrouverons les résultats de l'ACP sur le markdown.

II.3 Clustering avec 10 composantes principales



(a) Résultat kmeans (b) Résultat hclust

FIGURE 3 – Résultat clustering après ACP

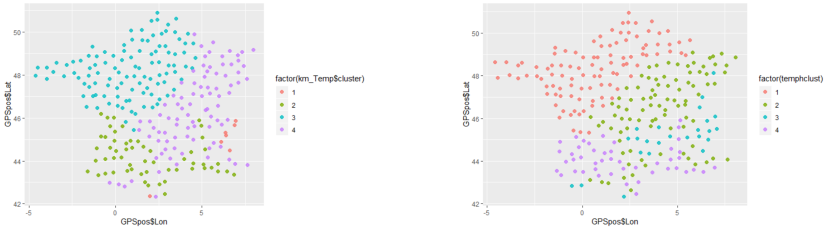
On constate qu'on perd quand même de l'information autour du point de coordonnées (0 ;47) en utilisant uniquement 10 composantes principales pour le clustering hierarchique.

Le kmeans lui renvoie des résultats plus ou moins similaires.

III Clustering en fonction des mesures de la température

Là aussi le choix de 4 cluster est justifié (voir le markdown).

III.1 Clustering avec kmeans et hclust sur les données initiales



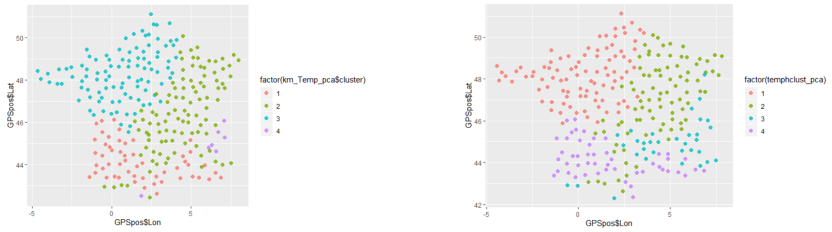
(a) Résultat kmeans (b) Résultat hclust

FIGURE 4 – Résultat clustering sur les mesures de température

Conclusion : Les résultats obtenus avec les deux méthodes sont quasiment les mêmes.

III.2 Analyse en composantes principales

On choisit de conserver de manière subjective 9 composantes qui correspondent à 90% de variances cumulées c'est à dire qu'on conserve 90% de l'information initiale car au dela de 9 composantes principales, la variance cumulée augmente très faiblement. Avec le choix imposé de 10 composantes principales nous avons accès à 91% de variances cumulées.



(a) Résultat kmeans après ACP (b) Résultat hclust après ACP

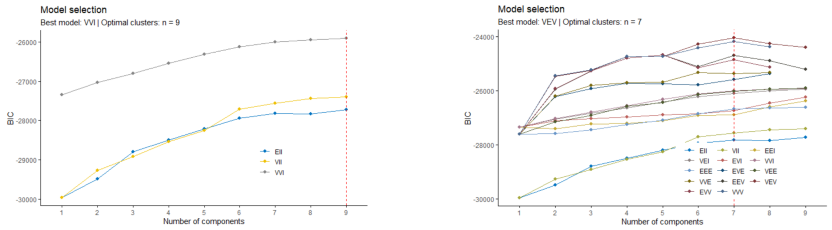
FIGURE 5 – Résultat clustering sur les mesures de température après ACP

Il n'y a pas de grande difference dans le regroupement des 259 sites.

III.3 Clustering basé les mélanges gaussiens

Les mélanges supposent que nos données sont des combinaisons des distributions gaussiennes. Nous avons obtenu le résultat suivant pour trois modèles estimés (EII, VII, VVI) dont la définition est donnée dans le tableau ci-dessous :

Modèle	distribution	volume	shape	orientation
EII	spherical	equal	equal	-
VII	spherical	variable	equal	-
VVI	diagonal	variable	variable	coordinate axis
VEV	ellipsoidal	variable	equal	variable



(a) Comparaison entre les 3 modèles (b) Meilleur modèle

FIGURE 6 – Selection du meilleur modèle

Le meilleur modèle est celui dont le BIC est le plus grand, c'est à dire VVI, on peut noté que pour les 3 modèles, le nombre de cluster optimal est de 9.

Comme quatrième modèle, nous avons décider de choisir celui dont le BIC est le maximum, c'est à dire le modèle VEV (ellipsoidal, equal volume) dont le nombre de cluster optimal est de 7.

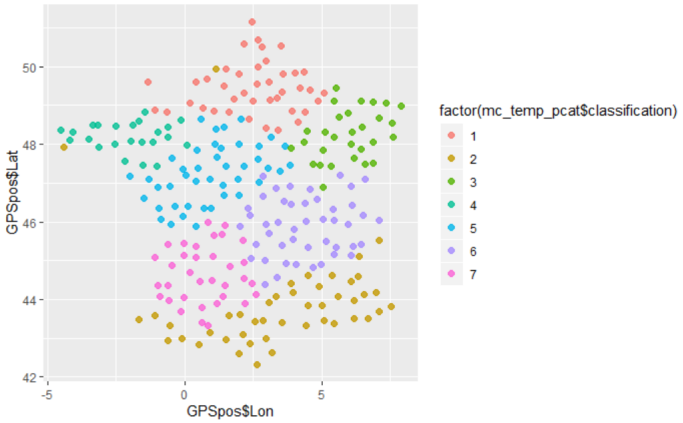


FIGURE 7 – Segmentation obtenu avec mclust (modèle VEV)

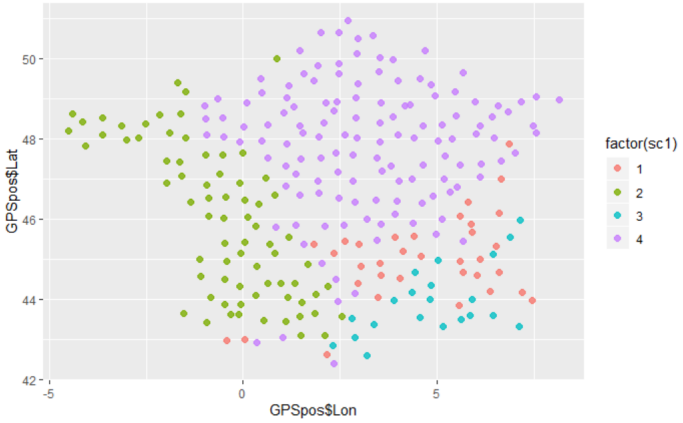


FIGURE 8 – Segmentation obtenu à l'aide de la fonction specc

IV Clustering en fonction des mesures de la temperature et du vent

On a décidé de concater les deux dataset et, de centrer et reduire le nouveau data set afin d'obtenir des données de même ordre car celles ci n'ont pas la même unité de mesure.
Ensuite ayant donc le nouveau data set dont les dimensions sont de (259x15720), nous avons décidé de faire une ACP afin de réduire la dimension et supprimé les rondondances au sein de ces données.
À l'aide de la méthode du coude, le choix de 6 clusters nous semble le plus approprié.
L'augmentation de la variance cumulée est faible à partir de 20 composantes principales et pour 21 composantes principales nous avons 90% de l'information initiale. Donc pour notre étude nous choisirons de garder 21 composantes principales. Nous décidons d'utiliser l'algorithme kmeans pour cette question.

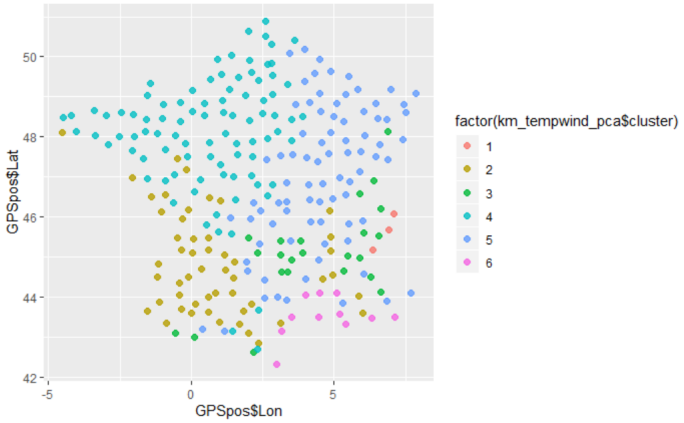


FIGURE 9 – Segmentation obtenu à l'aide de kmeans

Conclusion

A l'issue de ce projet, nous avons exploré quelques méthodes de clustering. Pour chacune des méthodes que nous avons utilisé d'abord sur les mesures de vent et de température, la segmentation pour chaque méthode donne pratiquement des résultats identiques.
En considérant à la fois sur la temperature et le vent, la segmentation finale obtenue semble cohérente par rapport au climat du territoire français.