Université de Nantes

Licence Professionnelle "métrologie chimique et nucléaire"

Module de statistiques

Méthodes factorielles - Exercices

Exercices : Classification ascendante hiérarchique

**A – Visualisation du regroupement des individus et de la construction du dendrogramme**

Dans un espace à deux dimensions, on considère les 6 points suivants

A 0,5 0,5

B 0,5 0,7

C 1 1

D 1 2

E 2 2

F 2 1,2

1/ Visualisez les points en les portant sur un graphique (référentiel orthonormé).

2/ Établissez la matrice des distances.

*Vous utiliserez comme distance la distance "city block".*

3/ A l'aide de la matrice des distances, identifiez les deux points les plus proches.

3/ Fusionnez les 2 points les plus proches et calculez les coordonnées du nouveau point.

*Le point résultant de la fusion s'appellera G.*

4/ Créez la première "arche" du dendrogramme

*Sa hauteur sera proportionnelle à la distance séparant les deux points avant leur fusion.*

5/ Mettez à jour la liste des points et le graphique.

*Rayez les points fusionnés, tant sur le graphique que dans la liste.*

6/ Répéter les étapes 2 à 5 tant que tous les points n'ont pas été fusionnés.

*NB : Plutôt que d'établir la matrice des distances à chaque itération, vous pouvez vous limiter au calcul des distances pour les couples de points qui semblent les plus proches.*

7/ Effectuez une coupe du dendrogramme et visualisez les classes sur le graphique initial.

6/ Ajustez, si nécessaire, le niveau de coupe du dendrogramme.

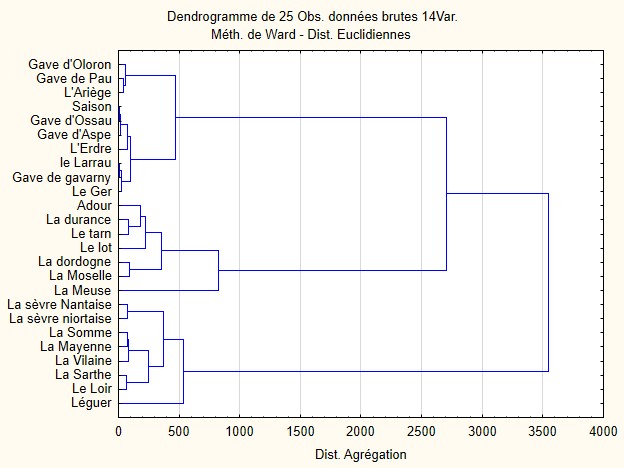
**B – Examen de Dendrogrammes et lien avec l'ACP**

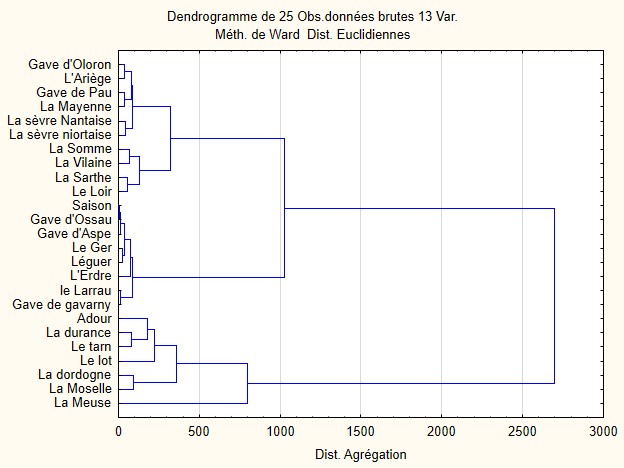
Les données ont été obtenues par Sébastien MATHIEU, promotion 2022-2023

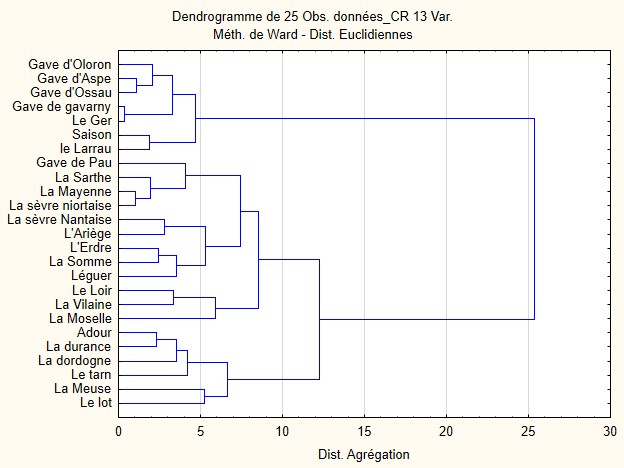
L'étudiant ayant obtenu le jeu de données a choisi de décrire des rivières au moyen de variables "géographiques" telles que la longueur du cours d'eau, son débit, le nombre de départements traversés ou le nombre de villes de plus de 2000 habitants mais également au moyen de variables "sportives" relatives à la pratique du kayak : classe de difficulté, nombre de club de kayak …

Les deux premiers dendrogrammes sont obtenus à partir des données brutes (= non centrées réduites), respectivement avec 14 puis 13 variables.

Le troisième dendrogramme est obtenu à partir des données centrées réduites, avec 13 variables.

**

**

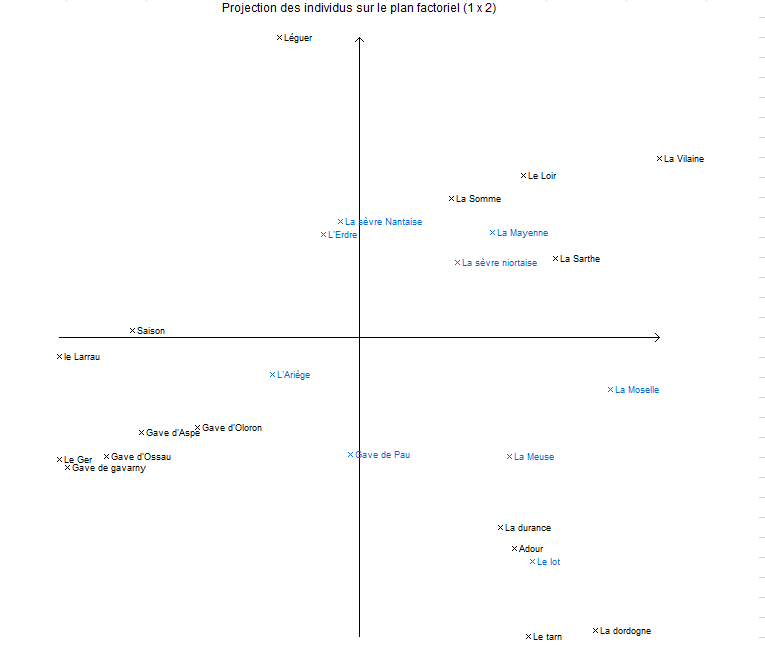
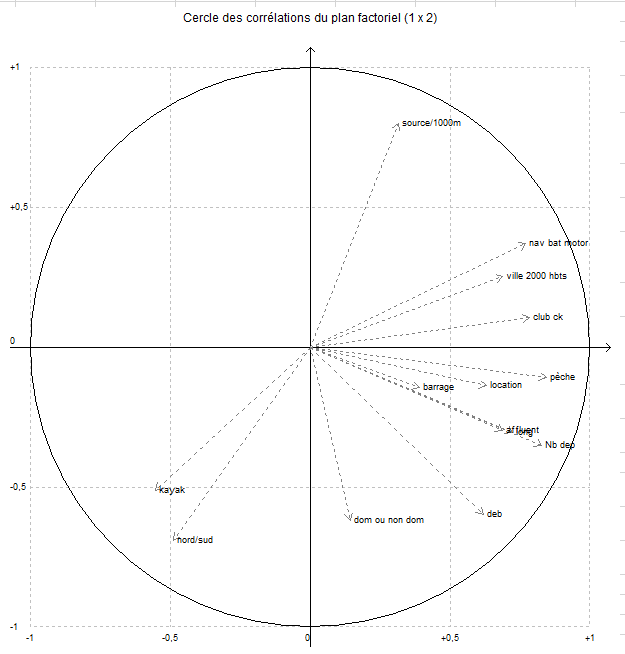
**

1/ Comparez les 3 dendrogrammes. En combien de classes paraît-il naturel de couper chacun d'entre eux ?

2/ Quels éléments peuvent expliquer l'aspect plus "ramassé" des graphes en données brutes ?

3/ Que peut-on dire du positionnement de l'Erdre sur les différents dendrogrammes ?

4/ Retrouve-t-on sur le premier plan principal les classes visibles sur le dendrogramme centré-réduit ?

 **