Sciences des données - TP2

Alexandre Theisse Louis-Vincent Capelli Tom Sartori

November 22, 2023

Question 1

Question 2

Question 3

Mesure de séparation des clusters

Voici les résultats obtenus pour les 2 mesures de séparation des clusters pour les différents nombres de clusters.

On constate que le score silhouette est maximal pour 2 clusters et diminue constamment par la suite. Même le score maximal (0.43) est plus proche de 0 que de 1, ce qui indique que les clusters sont assez mal séparés. Les scores correspondants aux grands nombres de clusters sont encore plus faibles (proches de 0.2), ce qui indique un fort chevauchement entre les clusters.

On constate également cette tendance à l'augmentation du chevauchement en observant les matrices d'overlaps. On remarque que les valeurs augmentent avec le nombre de clusters, ce qui indique que les clusters sont de plus en plus chevauchants. En effet, les meilleures valeurs d'overlaps obtenues entre 2 clusters sont supérieures à 1, ce qui indique que les clusters ne sont pas séparés et les moins bonnes valeurs sont de l'ordre de 3 ce qui indique un chevauchement important.

2 clusters

Silhouette score : 0.43

Overlaps 2 à 2 :

	C1	C2
C1	-	4.4
C2	-	-

3 clusters

Silhouette score : 0.34

Overlaps 2 à 2 :

	C1	C2	C3
C1	-	4.49	3.38
C2	-	-	1.74
C3	-	-	-

4 clusters

Silhouette score : 0.30

Overlaps 2 à 2 :

	C1	C2	C3	C4
C1	-	1.41	4.51	3.61
C2	-	-	1.09	3.34
C3	-	-	-	1.6
C4	-	-	-	-

5 clusters

Silhouette score : 0.26

Overlaps 2 à 2 :

	C1	C2	C3	C4	C5
C1	-	3.45	1.7	1.36	4.5
C2	-	-	1.02	3.36	1.47
C3	-	-	-	0.82	4.21
C4	-	-	-	-	0.98
C5	-	-	-	-	-

6 clusters

Silhouette score : 0.22

7 clusters

Silhouette score : 0.20

8 clusters

Silhouette score : 0.20

9 clusters

Silhouette score : 0.19

10 clusters

Silhouette score : 0.19

Cas particulier des 3 clusters

Question 4

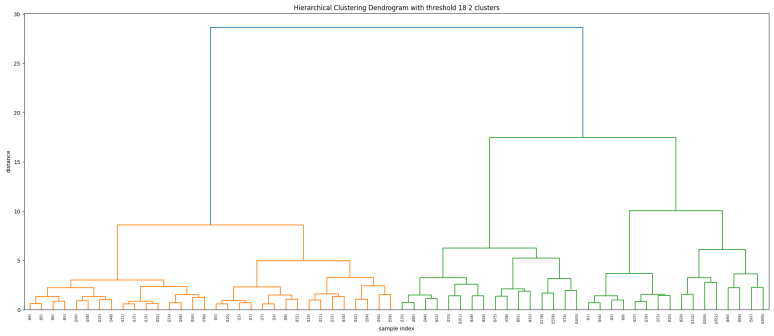
Différentes valeurs de seuil

En utilisant un clustering hierarchique avec la distance euclidienne et le lien de Ward, on obtient peut tester différentes valeurs de seuil pour obtenir un nombre de clusters donné. Voici les résultats obtenus pour les différentes valeurs de seuil testées.

NB : Les résultats sont aussi mauvais en utilisant un lien simple ou moyen.

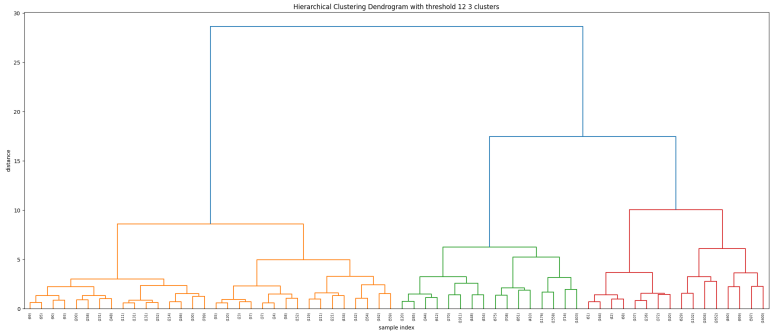
Seuil = 18

Pour un seuil de 18, on obtient 2 clusters.
Voici le dendrogramme obtenu :



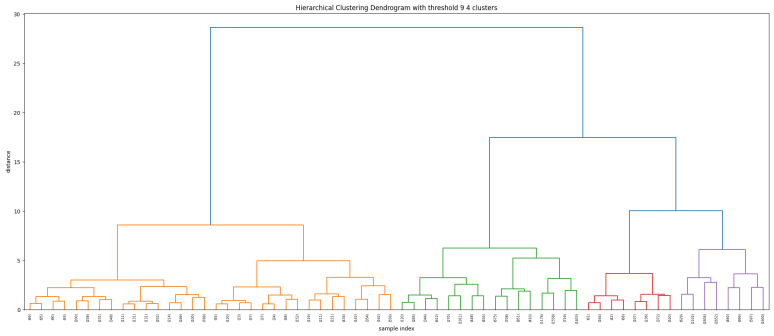
Seuil = 12

Pour un seuil de 12, on obtient 3 clusters.
Voici le dendrogramme obtenu :



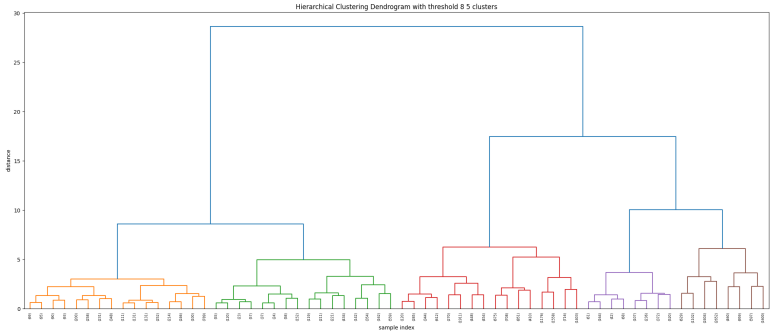
Seuil = 9

Pour un seuil de 9, on obtient 4 clusters.
Voici le dendrogramme obtenu :



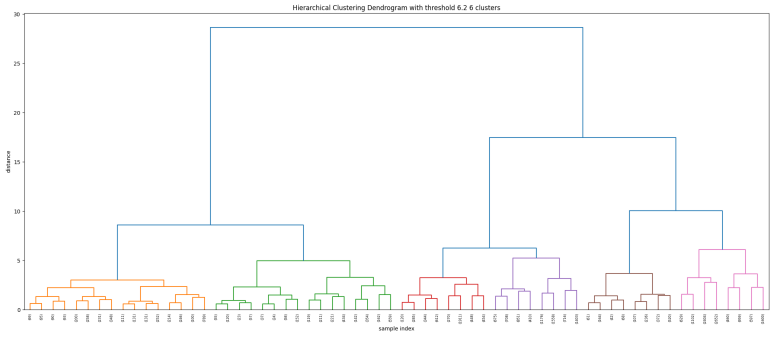
Seuil = 8

Pour un seuil de 8, on obtient 5 clusters.
Voici le dendrogramme obtenu :



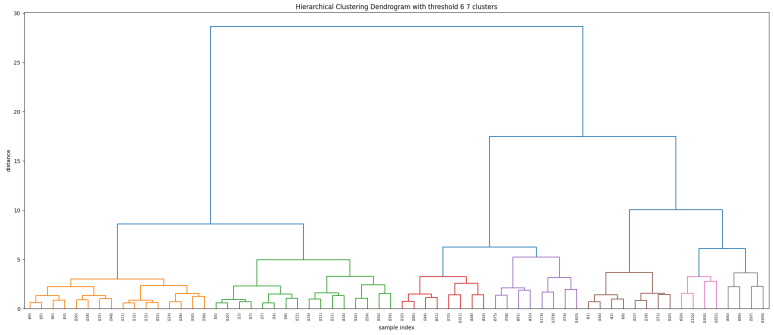
Seuil = 6.2

Pour un seuil de 6.2, on obtient 6 clusters.
Voici le dendrogramme obtenu :



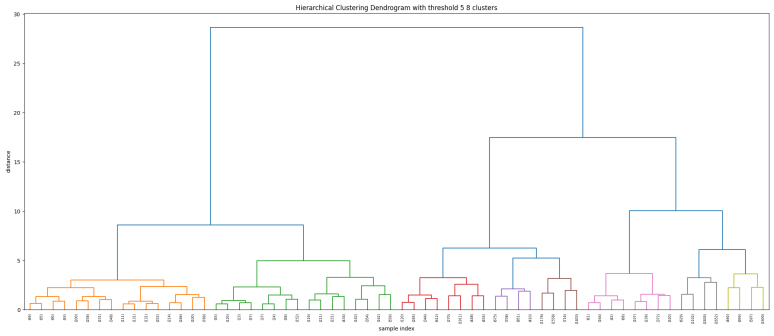
Seuil = 6

Pour un seuil de 6, on obtient 7 clusters.
Voici le dendrogramme obtenu :



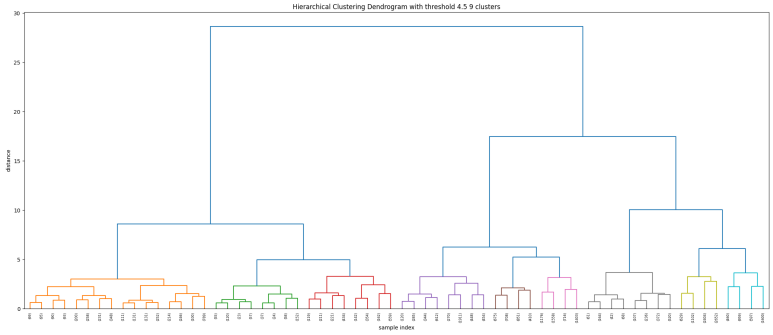
Seuil = 5

Pour un seuil de 5, on obtient 8 clusters.
Voici le dendrogramme obtenu :



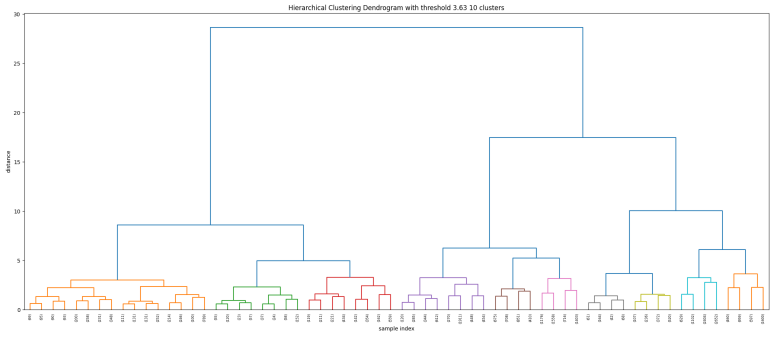
Seuil = 4.5

Pour un seuil de 4.5, on obtient 9 clusters.
Voici le dendrogramme obtenu :



Seuil = 3.63

Pour un seuil de 3.63, on obtient 10 clusters.
Voici le dendrogramme obtenu :



Mesure de séparation des clusters

Voici les résultats obtenus pour les 2 mesures de séparation des clusters pour les différents nombres de clusters.

On constate que le score silhouette est maximal pour 2 clusters et diminue constamment par la suite. Même le score maximal (0.45) est plus proche de 0 que de 1, ce qui indique que les clusters sont assez mal séparés. Les scores correspondants aux grands nombres de clusters sont encore plus faibles (proches de 0.2), ce qui indique un fort chevauchement entre les clusters.

On constate un overlap qui explose entre certains clusters. Par exemple, dans une configuration avec 2 clusters on obtient un overlap de 3667.46.

Seuil = 18

Pour un seuil de 18, on obtient 2 clusters. Le score silhouette est de 0.45.

Les overlaps 2 à 2 sont les suivants :

	C1	C2
C1	-	3667.46
C2	-	-

Seuil = 12

Pour un seuil de 12, on obtient 3 clusters. Le score silhouette est de 0.28.

Les overlaps 2 à 2 sont les suivants :

	C1	C2	C3
C1	-	3667.46	7802.60
C2	-	-	1.41
C3	-	-	-

Seuil = 9

Pour un seuil de 9, on obtient 4 clusters. Le score silhouette est de 0.25.

Les overlaps 2 à 2 sont les suivants :

	C1	C2	C3	C4
C1	-	3667.46	7802.60	954.16
C2	-	-	1.41	2.53
C3	-	-	-	4.21
C4	-	-	-	-

Seuil = 8

Pour un seuil de 8, on obtient 5 clusters. Le score silhouette est de 0.22.

Les overlaps 2 à 2 sont les suivants :

	C1	C2	C3	C4	C5
C1	-	1954.64	1695.18	7802.6	954.16
C2	-	-	1.08	2.19	1.60
C3	-	-	-	2.48	1.44
C4	-	-	-	-	4.21
C5	-	-	-	-	-

Seuil = 6.2

Pour un seuil de 6.2, on obtient 6 clusters. Le score silhouette est de 0.18.

Seuil = 6

Pour un seuil de 6, on obtient 7 clusters. Le score silhouette est de 0.15.

Seuil = 5

Pour un seuil de 5, on obtient 8 clusters. Le score silhouette est de 0.14.

Seuil = 4.5

Pour un seuil de 4.5, on obtient 9 clusters. Le score silhouette est de 0.13.

Seuil = 3.63

Pour un seuil de 3.63, on obtient 10 clusters. Le score silhouette est de 0.13.