

Classification Hiérarchique Ascendante Utilisation du logiciel R

Zakaria EL MOUMNAOUI

4 juillet 2020

Table des matières

1 Classification Hiérarchique Ascendante	5
1.1 Classification Hiérarchique Ascendante	5
1.1.1 Définitions	5
1.1.2 Visualisation des données	7
1.2 Algorithme de la classification	8
1.2.1 Préparation des données	8
1.2.2 Dissimilarité et matrice des distances	8
1.2.3 Fusion et choix du nombre de classes	19
1.2.4 Coupure du dendrogramme et interprétation des résultats	26
2 Application numérique avec logiciel R	30
Bibliographie	56

Introduction

Dès le début du $XX^{\text{ème}}$ siècle, le monde a vécu des changements révolutionnaires dans tous les domaines (économiques, sociaux, militaires...). Ces changements sont accompagnés d'une explosion de données, sous plusieurs formes. D'où le besoin d'outils sophistiqués pour manipuler ces données.

L'analyse des données est une discipline plus ou moins récente, ses bases sont connues depuis longtemps, mais elle n'a pu être développée qu'avec l'invention d'ordinateurs durant la $2^{\text{ème}}$ guerre mondiale, ce qui a rendu le traitement des grandes masses de données, faisable. L'analyse des données moderne était établie par les statisticiens Jean-Paul Benzecri, Chikio Hayashi et le psychiatre Louis Guttman, au $XX^{\text{ème}}$ siècle.

L'analyse des données est un champ scientifique multidimensionnel dirigé vers le traitement des données, afin d'extraire les informations qu'elles contiennent (Data Mining). Ainsi, on peut exploiter ces informations pour faire des prédictions et alors faire éventuellement des choix appropriés et prendre de bonnes décisions (Aide à la Décision).

Il existe plusieurs méthodes en analyse des données telles que l'analyse par réduction des dimensions et l'analyse par classification. Dans ce mémoire on va s'intéresser au $2^{\text{ème}}$ axe, l'analyse par classification et plus précisément la classification hiérarchique.

La classification hiérarchique est une méthode d'apprentissage non supervisé dans l'apprentissage automatique (Machine Learning). Les données à traiter sont à l'état brut, non modifiées et telles qu'elles existent à l'origine (pas de classes prédéfinies). Cette méthode est constituée de deux processus principaux, la classification hiérarchique ascendante (la plus utilisée) et la classification hiérarchique descendante. Ce sont deux algorithmes opposés l'un à l'autre dans la démarche du traitement des données.

Ce mémoire sera constitué de 2 chapitres. Un premier chapitre où l'on présentera la classification hiérarchique ascendante, un second chapitre concernera une application numérique de la méthode avec de grandes masses de données en utilisant le logiciel R.

Remerciements

Je souhaite tout d'abord remercier infiniment mon encadrante, le professeur Lalla Aicha Allamy, pour ses orientations judicieuses, ses précieux conseils, ses supports et ses encouragements tout au long de ce projet.

Je voudrais aussi remercier tous les membres de jury, le professeur Abdelaziz Nasroallah et le professeur Abdallah Mkhadri, qui ont bien voulu évaluer mon projet de fin d'études et faire partie du jury.

Je tiens également à remercier tous les enseignants de la Faculté des Sciences Semlalia et surtout les Professeurs du Département de Mathématiques.

Je n'oublie pas de remercier Allah en premier, mes parents qui ont toujours été présent pour me soutenir, mes proches amis et toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce projet.

Chapitre 1

Classification Hiérarchique Ascendante

1.1 Classification Hiérarchique Ascendante

1.1.1 Définitions

La classification hiérarchique est un algorithme qui regroupe les données dans des classes, suivant un critère bien choisi.

Il existe de nombreuses applications de la classification hiérarchique dans plusieurs domaines :

- 1) Biologie : règne animal, classification suivant l'ADN des être Humain.
- 2) Géographie : division géographique du Maroc.
- 3) Education : classification des étudiants dans une établissement scolaire.
- 4) Marketing et commerce : segmentation des profils des clients et recommandation des marchandises et des services (achat et location des voitures, produits alimentaires...), segmentation des posts de travail dans une société.
- 5) Divertissement : recommandation des multimédia (filmes, videos YouTube...).

Dans ce chapitre on va traiter la classification hiérarchique ascendante qui est la plus utilisée dans cette catégorie.

Définition 1 :

La classification hiérarchique ascendante est un algorithme qui consiste à considérer chaque donnée comme étant une classe au départ et essayer à

chaque itération de fusionner les classes qui sont proches entre elles jusqu'à les regrouper dans une seule classe, en se basant sur un critère bien choisi.

Remarque 1 :

- 1) *La classification s'intéresse à des tableaux de données individus-variables quantitatives.*
- 2) *Objectifs : production d'une structure (dendrogramme) permettant :*
 - *La mise en évidence de liens hiérarchique entre individus ou groupes d'individus.*
 - *La détection d'un nombre de classes "naturel" au sein de la population (Hidden patterns).*
- 3) *Le processus s'arrêtera automatiquement quand les données se regrouperont dans une seule classe, mais en prenant en considération l'étude à faire, on choisi une étape bien précise dans l'algorithme à considérer comme point d'arrêt.*

Pour plus de précision, on considère un ensemble fini Ω d'individus (données). On netra ω , un élément quelconque de Ω .
On suppose que l'on dispose d'une mesure de dissimilarité entre les classes.
Lorsque l'on parle de classification hiérarchique, on parle donc de l'existence d'une hiérarchie, que l'on notera H .

Définition 2 :

Une hiérarchie H est l'ensemble des classes (éléments de $P(\Omega)$), ensemble des parties de Ω) à toutes les étapes de l'algorithme, qui vérifie les propriétés suivantes :

- 1) $\emptyset \notin H$: aucune classe n'est vide.
- 2) $\Omega \in H$: au sommet de l'hiérarchie tous les individus sont groupés dans une seule classe.
- 3) $\forall \omega \in \Omega, \{\omega\} \in H$: en bas de l'hiérarchie, tous les individus se trouvent seuls (une classe par individus).
- 4) $\forall (h_1, h_2) \in H^2, h_1 \cap h_2 = \emptyset$ ou $h_1 \subset h_2$ ou $h_2 \subset h_1$: si l'on considère deux classes du regroupement, soit elles sont disjointes, soit l'une est incluse dans l'autre.

Pour illustrer ceci, on présente un exemple.

Exemple 1 :

Soit $\Omega = \{A, B, C, D, E, F, G, H, I, J, K\}$ un ensemble de points. Une hiérarchie de Ω peut être comme suit :

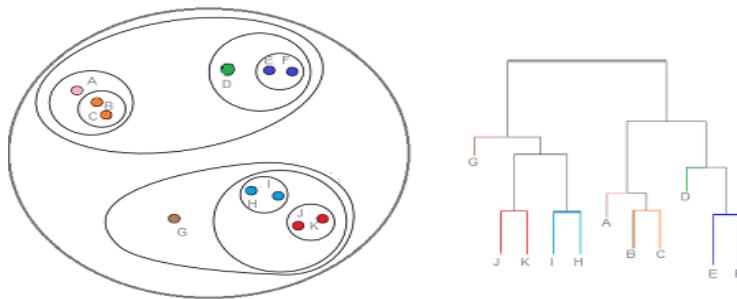


FIGURE 1.1 – Exemple d’une hiérarchie de parties de Ω

1.1.2 Visualisation des données

La visualisation des données se fait à travers un graphique typique appelé «dendrogramme». Un dendrogramme est un diagramme sous forme d’un arbre, sur l’axe des abscisses figurent les données initiales et sur l’axe des ordonnées une échelle est établie pour mesurer les dissimilarités ou les indices d’agrégation entre les classes.

La visualisation par dendrogramme est une technique visant à partitionner une population en différentes classes ou sous-groupes.

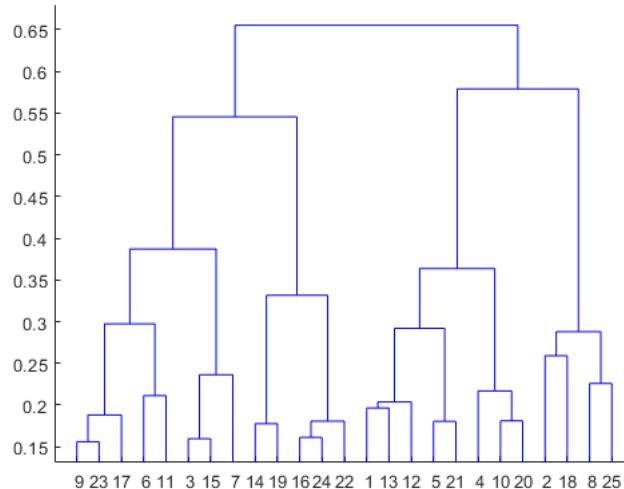


FIGURE 1.2 – Exemple de dendrogramme

1.2 Algorithme de la classification

Dans cet algorithme on cherche à ce que les individus regroupés au sein d'une même classe soient les plus semblables possibles (homogénéité intra-classe), tandis que les classes soient le plus dissemblables (hétérogénéité inter-classe).

L'algorithme est basé sur les points suivants :

1. Préparation des données
2. Critère de dissimilarité et matrice des distances
3. Fusion et choix du nombre de classes.

1.2.1 Préparation des données

La préparation des données est la première tâche à faire, en important les données existantes ou bien en rentrant les données directement.

Remarque 2 :

On est amené des fois à centrer et/ou réduire les données. On peut aussi reconstruire le problème des données manquantes et alors soit on les supprime ou bien on les estime.

1.2.2 Dissimilité et matrice des distances

Soit E un sous-ensemble de \mathbb{R}^p de cardinal n et soient, à une étape t_m de l'algorithme, les m classes de données de $P(E)$ suivantes :

$C_1 = \{p_{11}, \dots, p_{1r_1}\}, \dots, C_m = \{p_{m1}, \dots, p_{mr_m}\}$ et d une distance sur \mathbb{R}^p (par exemple la distance euclidienne).

Définition 3 :

La dissimilité est un critère de comparaison entre les classes de données, notée $dissim(C_i, C_j)$, C_i et C_j sont deux classes de la hiérarchie H à construire.

Définition 4 :

La matrice des distances est une matrice dont les coefficients sont les valeurs des dissimilarités entre les classes deux à deux.

On écrit l'algorithme de la classification hiérarchique ascendante, comme suit :

Etant donnés un ensemble $E = \{p_1, \dots, p_n\}$ et un critère de dissimilarité "dissm".

```

for (i = 1 to n)
     $C_i = \{p_i\}$ 
end
P = { $C_1, \dots, C_n$ }
while P.size > 1 do
{
    ( $C_{min1}, C_{min2}$ ) = minimum dissim( $C_i, C_j$ ) for all  $C_i, C_j$  in P
    add { $C_{min1}, C_{min2}$ } to P
    delete  $C_{min1}$  and  $C_{min2}$  from P
}
end

```

Remarque 3 :

1. La dissimilarité dépend de la distance choisie.
2. Les deux classes qui ont la dissimilarité la plus faible entre elles vont être fusionnées.
3. La matrice des distances change à chaque étape du processus de regroupement des classes, suivant le critère était choisi.

On présente dans la suite quelques critères usuels de dissimilité.

Soient les m classes fixées ci-dessus, C_1, \dots, C_m de cardinal r_1, \dots, r_m , respectivement.

1) Critère du minimum ou lien simple :

On considère le minimum des distances entre les classes deux à deux :

$$\forall 1 \leq i \neq j \leq m, \quad dissim(C_i, C_j) = \min_{\substack{1 \leq k \leq r_i \\ 1 \leq l \leq r_j}} (d(p_{ik}, p_{jl})).$$

On illustre ceci par la figure suivante :



Chaque critère a des avantages et des inconvénients. Pour ce critère on cite :

- Avantages : Ce critère permet de séparer les classes qui sont loin entre elles.

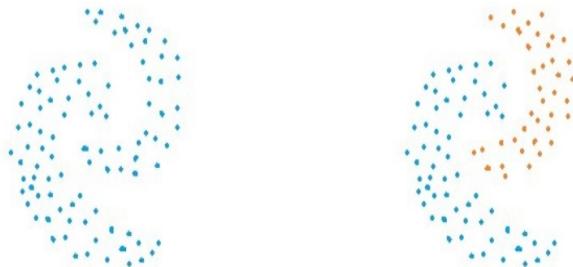


FIGURE 1.3 – Données non-elliptiques avec écart. Données réelles à gauche contre données classifiées à droite



FIGURE 1.4 – Données elliptiques avec écart. Données réelles à gauche contre données classifiées à droite

- Inconvénients : Ce critère ne peut pas séparer les données qui sont chevauchées (effet de chaîne).



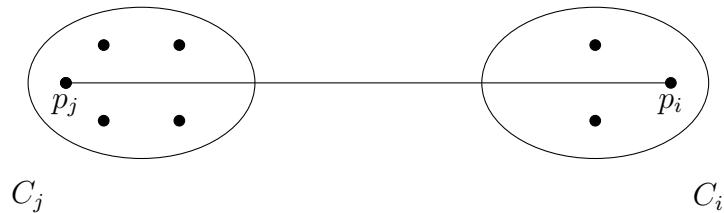
FIGURE 1.5 – Données réelles à gauche contre données classifiées à droite

2) Critère du maximum ou lien complet :

On considère le maximum des distances entre les classes deux à deux :

$$\forall 1 \leq i \neq j \leq m, \quad dissim(C_i, C_j) = \max_{\substack{1 \leq k \leq r_i \\ 1 \leq l \leq r_j}} (d(p_{i_k}, p_{j_l})).$$

On illustre ceci par la figure suivante :



- Avantages : Ce critère permet de séparer les classes qui sont proches entre elles.



FIGURE 1.6 – Données réelles à gauche contre données classifiées à droite

- Inconvénients : Ce critère est biaisé vers les grosses classes, c'est à dire il classe les données de manière à ce que les petites classes dominent des données de grosses classes.



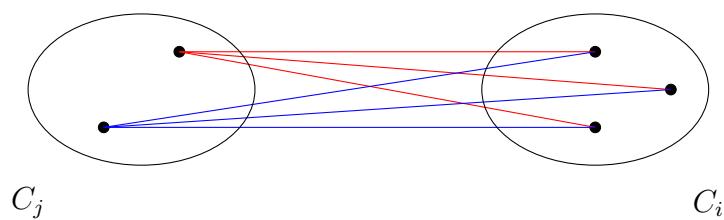
FIGURE 1.7 – Données réelles à gauche contre données classifiées à droite

3) Critère de la moyenne :

On considère la moyenne des distances entre les classes deux à deux :

$$\forall 1 \leq i \neq j \leq m, \quad dissim(C_i, C_j) = \frac{1}{r_i \times r_j} \sum_{1 \leq k \leq r_i} \sum_{1 \leq l \leq r_j} (d(p_{ik}, p_{jl})).$$

On illustre ceci par la figure suivante :



- Avantages : Ce critère permet de séparer les classes qui sont proches entre elles.
- Inconvénients : Ce critère est biaisé vers les grosses classes, de plus elle est coûteuse au nombre d'opérations à effectuer.

Définition 5 :

Soit $E = \{p_1, \dots, p_n\}$ un ensemble de \mathbb{R}^p .

Et $P_m = (C_1 = \{p_{1_1}, \dots, p_{1_{r_1}}\}, \dots, C_m = \{p_{m_1}, \dots, p_{m_{r_m}}\})$ une partition de E .

- L'inertie totale de E est $I_T = \frac{1}{n} \sum_{i=1}^n d^2(p_i, g)$.

- L'inertie intra-classe de P_m est la somme des inerties totales des classes C_j de P , $j = 1, \dots, m$:

$$I_W = \frac{1}{n} \sum_{j=1}^m \sum_{i=1}^{r_j} d^2(p_{ji}, g_j)$$

- L'inertie inter-classe est :

$$I_B = \frac{1}{n} \sum_{j=1}^m r_j d^2(g_j, g).$$

Avec g le barycentre de E , g_j le barycentre de C_j , $j = 1, \dots, m$ et d une distance sur l'espace \mathbb{R}^p .

Le résultat suivant est d'une importance dans la décomposition d'inertie.

Théorème 1 (Décomposition de Huygens) :

Sous les hypothèses de la définition 5 on a : $I_T = I_W + I_B$.

On peut voir la décomposition en 2D dans la figure suivante :

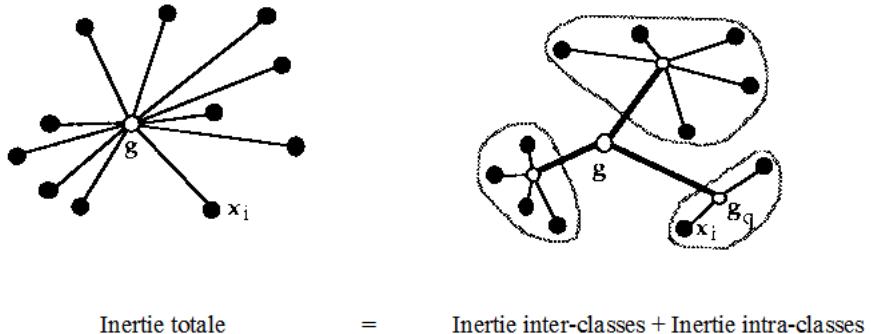


FIGURE 1.8 – Décomposition d'inertie

Preuve 1 : (Décomposition de Huygens)

Soit $E = \{p_1, \dots, p_n\}$ un ensemble de \mathbb{R}^p .

Soit $P_m = (C_1 = \{p_{11}, \dots, p_{1r_1}\}, \dots, C_m = \{p_{m1}, \dots, p_{mr_m}\})$ une partition de E , on a :

$$\begin{aligned}
I_T &= \frac{1}{n} \sum_{i=1}^n d^2(p_i, g). \\
&= \frac{1}{n} \sum_{i=1}^n \|p_i - g\|^2, \quad \mathbb{R}^p \text{ est un espace euclidien.} \\
&= \frac{1}{n} \sum_{k=1}^m \sum_{p_j \in C_k} \|p_j - g\|^2, \quad \text{somme par paquets disjoints.} \\
&= \frac{1}{n} \sum_{k=1}^m \sum_{p_i \in C_k} \|(p_j - g_k) + (g_k - g)\|^2.
\end{aligned}$$

Où $C_1 \cup C_2 \dots \cup C_m = \Omega$ et $C_r \cap C_t = \emptyset$, $r \neq t$.

$$\begin{aligned} I_T &= \frac{1}{n} \sum_{k=1}^m \sum_{p_j \in C_k} \left[\|p_j - g_k\|^2 + 2 \langle p_j - g_k, g_k - g \rangle + \|g_k - g\|^2 \right]. \\ &= \frac{1}{n} \sum_{k=1}^m \sum_{p_j \in C_k} \|p_j - g_k\|^2 + \frac{1}{n} \sum_{k=1}^m \sum_{p_j \in C_k} \|g_k - g\|^2. \end{aligned}$$

$$Car \frac{2}{n} < \sum_{k=1}^m \sum_{p_j \in C_k} p_j - \sum_{k=1}^m \sum_{p_j \in C_k} g_k, \sum_{k=1}^m \sum_{p_j \in C_k} (g_k - g) > = 0.$$

Il vient du fait que $\sum_{k=1}^m \sum_{p_j \in C_k} p_j = \sum_{k=1}^m \sum_{p_j \in C_k} g_k$. car $g_k = \frac{1}{|C_k|} \sum_{p_j \in C_k} p_j$.

$$D'où I_T = I_W + I_B.$$

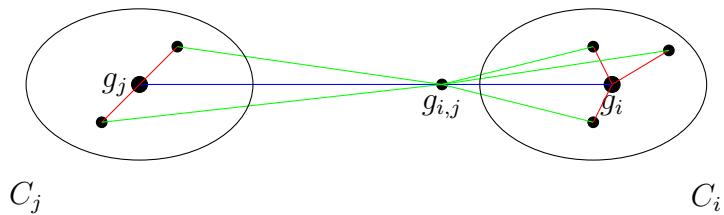
4) Critère de Ward :

A chaque étape on regroupe les deux classes dont leur agrégation produit une diminution de l'inertie inter-classe minimale.

$$\forall 1 \leq i \neq j \leq m, \quad dissim(C_i, C_j) = \frac{r_i \times r_j}{r_i + r_j} d^2(g_i, g_j) = I_{C_i \cup C_j} - (I_{C_i} + I_{C_j}).$$

Avec g_i est le barycentre de C_i , g_j est le barycentre de C_j , $I_{C_i \cup C_j}$ est l'inertie totale de $C_i \cup C_j$, I_{C_i} est l'inertie totale de C_i , I_{C_j} est l'inertie totale de C_j et d^2 est la distance euclidienne au carré.

On illustre dans \mathbb{R}^2 cette méthode par la figure suivante :



Avec $g_{i,j}$ est le barycentre de $C_j \cup C_i$.

- Avantages : Ce critère permet de séparer les classes qui sont proches entre elles et il est performant dans le cas d'effet de chaîne (données chevauchées).
- Inconvénients : Ce critère est biaisé vers les grosses classes et il est sensible aux données aberrantes (extrêmes).

Remarque 4 :

La nature des données et le choix du critère de dissimilarité influencent la matrice des distances et donc la classification des données, comme on le voit dans la figure ci-dessous.

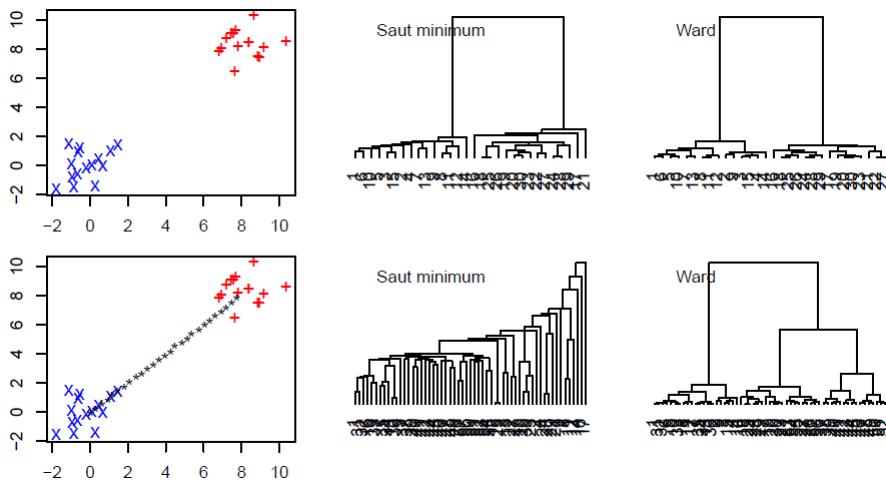


FIGURE 1.9 – Exemple d’effet de chaîne

- Commentaire sur la FIGURE 1.9 :

Les données (figures à gauche) sont réparties en deux groupes, un groupe représenté par des («+») et l’autre représenté par des («x»). De plus, sur la figure du bas, un bruit est ajouté («*»). En utilisant le critère du minimum, qui est sensible au bruit, on ne retrouve pas les groupes. Par contre, avec le critère de Ward, même avec le bruit, les deux groupes peuvent être distingués.

Maintenant on pense à améliorer notre hiérarchie et on se pose la question suivante :

Quand une partition est-elle dite bonne ?

Réponse :

- 1) Si les individus d’une même classe sont proches.
- 2) Si les individus de deux classes différentes sont éloignés.

La décomposition de Huygens mesure cette similarité entre les individus et entre les classes en se basant sur le changement de l’inertie au cours de l’algorithme et plus précisément, elle nous permet de suggérer un indicateur de qualité de la partition à chaque étape.

Cet indicateur est défini comme suit :

Définition 6 :

On appelle indicateur de qualité d'une partition à une étape donnée, la quantité :

$$R^2 = \frac{I_B}{I_T}.$$

Remarque 5 :

On a $0 \leq R^2 \leq 1$, plus R^2 est proche de 1, plus la partition est meilleure.

- 1) $R^2 = 0 \implies \forall j = 1, \dots, m, g_j = g$: les classes ont la même moyenne, on ne peut donc les classifier.
- 2) $R^2 = 1 \implies \forall j = 1, \dots, m$ et $i = 1, \dots, j$, $p_{ji} = g_j$: les individus d'une même classe sont identiques. Donc les classes sont très homogènes (i.e ceci est l'idéal pour classifier).

-Attention :

Ce critère ne peut être jugé comme absolu car il dépend du nombre d'individus et du nombre de classes, il permet juste de comparer deux partitions d'un ensemble E, de même nombre de classes.

On dispose aussi d'un autre coefficient de qualité d'une partition.

Définition 7 :

On appelle indice de silhouette d'un individu p_j de la classe C_j noté $s(p_j)$, la quantité :

$$s(p_j) = \frac{b(p_j) - a(p_j)}{\max(a(p_j), b(p_j))}.$$

Où $a(p_j) = \frac{1}{|C_j|-1} \sum_{p \in C_j, p \neq p_j} d(p_j, p)$ est la distance moyenne du point p_j à son groupe C_j et $b(p_j) = \min_{k \neq j} \frac{1}{|C_k|} \sum_{p \in C_k} d(p_j, p)$ est la distance moyenne du point p_j à son groupe voisin.

Remarque 6 :

- 1) $-1 \leq s(p_j) \leq 1$.
- 2) Une valeur de $s(p_j)$ proche de 1 signifie que le point p_j est cohérent avec sa classe mère C_j , une valeur nulle signifie que le point p_j est sur la frontière des deux classes (classe mère et classe voisine) et une valeur proche de -1

signifie que le point est cohérent avec la classe voisine plus que la classe mère.

3) On peut comparer l'homogénéité des groupes dans leur partition, en examinant la moyenne des indices de silhouette dans chaque groupe de cette partition moyennant la quantité :

$$S(C_k) = \frac{1}{|C_k|} \sum_{p \in C_k} s(p).$$

Les groupes ayant les coefficients de silhouette les plus forts sont les plus homogènes.

4) Sur l'ensemble de la classification, l'indice de silhouette est donné par :

$$S = \frac{1}{m} \sum_{k=1}^m \frac{1}{|C_k|} \sum_{p \in C_k} s(p).$$

On illustre la méthode par la figure suivante :

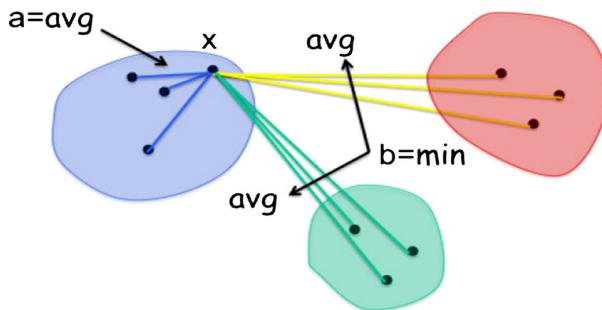


FIGURE 1.10 – Calcul de l'indice de silhouette du point x

Remarque 7 :

1) L'inertie totale étant constante, on essaie de minimiser la perte d'inertie inter-classe qui ne cesse que de diminuer, ce qui revient à minimiser le gain d'inertie intra-classe qui augmente, pour aboutir à un choix optimal et donc à une bonne classification.

2) Il existe plusieurs critères de dissimilarité autres que ceux définis auparavant. Donc selon la nature des données on essaie de choisir le plus approprié.

A ce stade, on se pose la question suivante : "quand le processus doit-il s'arrêter" ?

La réponse à cette question est le but de la section suivante.

1.2.3 Fusion et choix du nombre de classes

À chaque étape de l'algorithme on fusionne deux classes qui ont la dissimilarité minimale parmi les autres, dans une même nouvelle classe, et les autres classes de la hiérarchie H restent invariantes à priori jusqu'à l'étape suivante.

Ce processus de fusion se termine automatiquement par le regroupement des données dans une seule classe. Le choix du nombre de classes est un problème fondamental. Il n'existe pas de méthode générale pour le résoudre. Soit on a déjà le nombre de classes évidant (à partir de la nature des données), ou bien on le choisit en se basant sur le graphe de gain d'inertie intra-classe (i.e la perte d'inertie inter-classe) par la méthode du coude. Cette méthode consiste à choisir le nombre de classes en se basant sur la déviation aiguë dans la courbe.

Dans la figure suivante on voit deux déviations, il est toujours évident de choisir deux classes, mais on essaie de prendre d'autres, comme on le voit sur Figure 1.11. Il existe 3 classes à choisir.

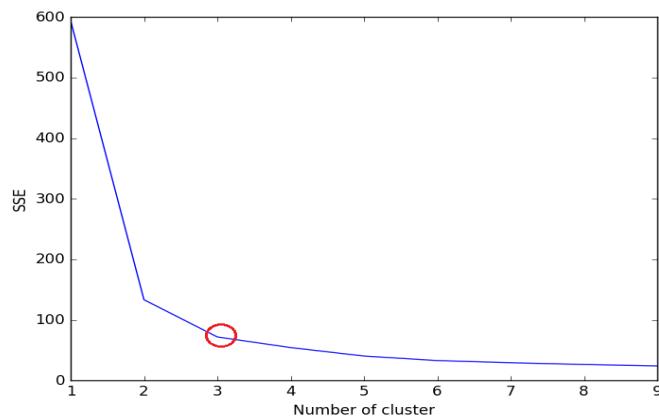


FIGURE 1.11 – Méthode du coude

Pour illustrer ces notions et voir les différentes méthodes utilisées dans l'algorithme, on considère l'exemple suivant :

Exemple 2 :

On considère le tableau de données, suivant :

	x	y
p_1	0.4	0.53
p_2	0.22	0.38
p_3	0.35	0.32
p_4	0.26	0.19
p_5	0.08	0.41
p_6	0.45	0.3

Pour obtenir les classes, on utilisera, respectivement le critère du minimum, le critère du maximum et le critère de la moyenne.

Remarque 8 :

- 1) Pour cet exemple on n'a pas besoin de la préparation des données.
- 2) Pour ces trois critères, on déroulera l'algorithme sans imposer de point d'arrêt. L'algorithme (ou le processus) s'arrêtera alors, une fois que toutes les données seront regroupées dans une seule classe.

a) Critère du minimum :

1^{ère} étape : On calcule la matrice des distances en utilisant la distance euclidienne et on obtient :

	p_1	p_2	p_3	p_4	p_5	p_6
p_1	0					
p_2	0.23	0				
p_3	0.22	0.15	0			
p_4	0.37	0.2	0.15	0		
p_5	0.34	0.14	0.28	0.29	0	
p_6	0.23	0.25	0.11	0.22	0.39	0

La valeur 0.11 est le minimum des valeurs, alors p_3 et p_6 vont être fusionnées dans une classe commune $\{p_3, p_6\}$.

2^{ème} étape : On recalcule la matrice de distances à nouveau avec les nouvelles classes.

	p_1	p_2	$\{p_3, p_6\}$	p_4	p_5
p_1	0				
p_2	0.23	0			
$\{p_3, p_6\}$	0.22	0.15	0		
p_4	0.37	0.2	0.15	0	
p_5	0.34	0.14	0.28	0.29	0

La valeur 0.14 est le minimum des valeurs alors p_2 et p_5 vont être fusionnées dans une classe commune $\{p_2, p_5\}$.

3^{ème} étape : On recalcule à nouveau la matrice des distances.

	p_1	$\{p_2, p_5\}$	$\{p_3, p_6\}$	p_4
p_1	0			
$\{p_2, p_5\}$	0.23	0		
$\{p_3, p_6\}$	0.22	0.15	0	
p_4	0.37	0.2	0.15	0

Remarque 9 :

La valeur 0.15 est le minimum mais elle figure deux fois dans la matrice. Là on choisit la première dans la matrice et donc les classes $\{p_2, p_5\}$ et $\{p_3, p_6\}$ vont être fusionnées dans une classe commune $\{p_2, p_5, p_3, p_6\}$.

4^{ème} étape : Après avoir recalculé la matrice des distances, on obtient

	p_1	$\{p_2, p_5, p_3, p_6\}$	p_4
p_1	0		
$\{p_2, p_5, p_3, p_6\}$	0.22	0	
p_4	0.37	0.15	0

La valeur 0.15 est le minimum donc $\{p_2, p_5, p_3, p_6\}$ et p_4 vont être fusionnées dans une classe commune $\{p_2, p_5, p_3, p_6, p_4\}$.

5^{ème} étape : On recalcule à nouveau la matrice des distances.

	p_1	$\{p_2, p_5, p_3, p_6, \}$
p_1	0	
$\{p_2, p_5, p_3, p_6, p_4\}$	0.22	0

La dernière valeur est 0.22 donc $\{p_2, p_5, p_3, p_6, p_4\}$ et p_1 vont être fusionnées dans une classe commune $\{p_2, p_5, p_3, p_6, p_4, p_1\}$ qui regroupe toutes les données.

On obtient alors la représentation graphique des résultats avec le dendrogramme suivant :

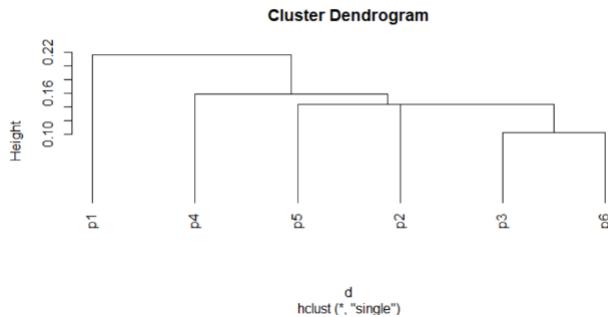


FIGURE 1.12 – Dendrogramme relatif au critère du minimum

b) Critère du maximum :

1^{ère} étape : Pour la première étape, la matrice des distances est la même.

	p_1	p_2	p_3	p_4	p_5	p_6
p_1	0					
p_2	0.23	0				
p_3	0.22	0.15	0			
p_4	0.37	0.2	0.15	0		
p_5	0.34	0.14	0.28	0.29	0	
p_6	0.23	0.25	0.11	0.22	0.39	0

La valeur 0.11 est le minimum des valeurs, alors p_3 et p_6 vont être fusionnées dans une classe commune $\{p_3, p_6\}$.

2^{ème} étape : On recalcule à nouveau la matrice des distances avec les nouvelles classes.

	p_1	p_2	$\{p_3, p_6\}$	p_4	p_5
p_1	0				
p_2	0.23	0			
$\{p_3, p_6\}$	0.23	0.25	0		
p_4	0.37	0.2	0.22	0	
p_5	0.34	0.14	0.39	0.29	0

La valeur 0.14 est le minimum des valeurs alors p_2 et p_5 vont être fusionnées dans une classe commune $\{p_2, p_5\}$

3^{ème} étape : On recalcule à nouveau la matrice.

	p_1	$\{p_2, p_5\}$	$\{p_3, p_6\}$	p_4
p_1	0			
$\{p_2, p_5\}$	0.34	0		
$\{p_3, p_6\}$	0.22	0.15	0	
p_4	0.37	0.29	0.22	0

La valeur 0.22 est le minimum donc $\{p_3, p_6\}$ et p_4 vont être fusionnées dans une classe commune $\{p_3, p_6, p_4\}$.

4^{ème} étape : On recalcule à nouveau la matrice.

	p_1	$\{p_2, p_5\}$	$\{p_3, p_6, p_4\}$
p_1	0		
$\{p_2, p_5\}$	0.34	0	
$\{p_3, p_6, p_4\}$	0.37	0.39	0

La valeur 0.34 est le minimum donc $\{p_2, p_5\}$ et p_1 vont être fusionnées dans une classe commune $\{p_2, p_5, p_1\}$.

5^{ème} étape : On recalcule à nouveau la matrice.

	$\{p_2, p_5, p_1\}$	$\{p_3, p_6, p_4\}$
$\{p_2, p_5, p_1\}$	0	
$\{p_3, p_6, p_4\}$	0.39	0

La dernière valeur est 0.39 donc $\{p_2, p_5, p_1\}$ et $\{p_3, p_6, p_4\}$ vont être fusionnées dans une classe commune $\{p_2, p_5, p_1, p_3, p_6, p_4\}$ qui regroupe toutes les données.

On obtient la représentation graphique des résultats avec le dendrogramme suivant :

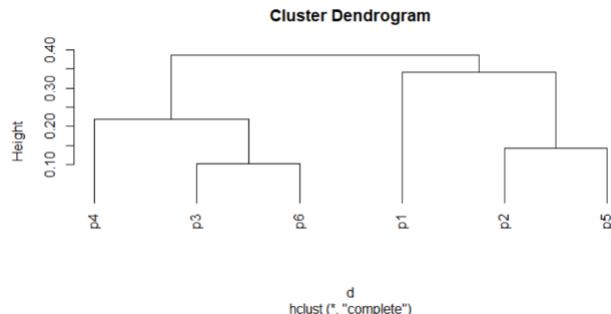


FIGURE 1.13 – Dendrogramme relatif au critère du maximum

c) Critère de la moyenne :

1^{ère} étape : Pour la première étape, la matrice des distances est la même qu'auparavant.

	p_1	p_2	p_3	p_4	p_5	p_6
p_1	0					
p_2	0.23	0				
p_3	0.22	0.15	0			
p_4	0.37	0.2	0.15	0		
p_5	0.34	0.14	0.28	0.29	0	
p_6	0.23	0.25	0.11	0.22	0.39	0

La valeur 0.11 est le minimum des valeurs, alors p_3 et p_6 vont être fusionnées dans une classe commune $\{p_3, p_6\}$.

2^{ème} étape : On recalcule la matrice de distances à nouveau avec les nouvelles classes.

	p_1	p_2	$\{p_3, p_6\}$	p_4	p_5
p_1	0				
p_2	0.23	0			
$\{p_3, p_6\}$	0.23	0.2	0		
p_4	0.37	0.2	0.19	0	
p_5	0.34	0.14	0.34	0.29	0

La valeur 0.14 est le minimum des valeurs alors p_2 et p_5 vont être fusionnées dans une classe commune $\{p_2, p_5\}$.

3^{ème} étape : On recalcule à nouveau la matrice.

	p_1	$\{p_2, p_5\}$	$\{p_3, p_6\}$	p_4
p_1	0			
$\{p_2, p_5\}$	0.29	0		
$\{p_3, p_6\}$	0.23	0.27	0	
p_4	0.37	0.25	0.19	0

La valeur 0.19 est le minimum donc $\{p_3, p_6\}$ et p_4 vont être fusionnées dans une classe commune $\{p_3, p_6, p_4\}$.

4^{ème} étape : On recalcule à nouveau la matrice.

	p_1	$\{p_2, p_5\}$	$\{p_3, p_6, p_4\}$
p_1	0		
$\{p_2, p_5\}$	0.29	0	
$\{p_3, p_6, p_4\}$	0.3	0.26	0

La valeur 0.26 est le minimum donc $\{p_2, p_5\}$ et $\{p_3, p_6, p_4\}$ vont être fusionnées dans une classe commune $\{p_2, p_5, p_3, p_6, p_4\}$.

5^{ème} étape : On recalcule à nouveau la matrice.

	p_1	$\{p_2, p_5, p_3, p_6, p_4\}$
p_1	0	
$\{p_2, p_5, p_3, p_6, p_4\}$	0.3	0

La dernière valeur est 0.3 donc $\{p_2, p_5, p_3, p_6, p_4\}$ et p_1 vont être fusionnées dans une classe commune $\{p_1, p_2, p_5, p_3, p_6, p_4\}$ qui regroupe toutes les données.

On obtient la représentation graphique des résultats avec le dendrogramme suivant :

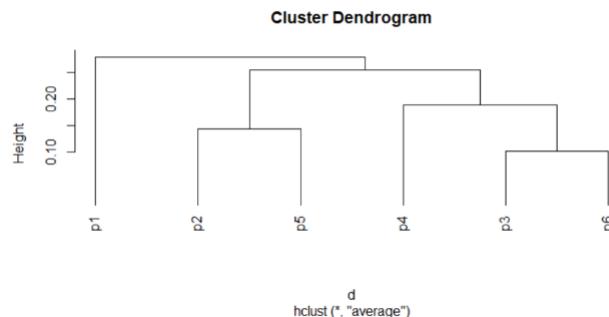


FIGURE 1.14 – Dendrogramme relatif au critère de la moyenne

Remarque 10 :

Pour les trois critères, les valeurs des dissimilarités augmentent d'une étape à l'autre, et ceci vient du fait que les classes de données deviennent de plus en plus dissimilaires entre elles d'une étape à l'autre.

Dans le paragraphe suivant on s'intéressera à la coupure d'un dendrogramme et à l'interprétation des résultats obtenus.

1.2.4 Coupure du dendrogramme et interprétation des résultats

Dans la première partie de ce paragraphe, on essaie de visualiser les résultats et de préciser les classes à considérer. La coupure du dendrogramme est un moyen d'effectuer cette tâche. On considère un segment horizontal qui coupe la hiérarchie H en des points particuliers. Chaque point de la coupure correspond à une classe de la hiérarchie H, le niveau de placement du segment de la coupure donne à priori un nombre de classes différent. En définissant un niveau de la coupure, on définit une partition et vice-versa.

Exemple 3 :

On présente le deuxième dendrogramme initial de l'exemple précédent que l'on coupe après en deux niveaux différents :

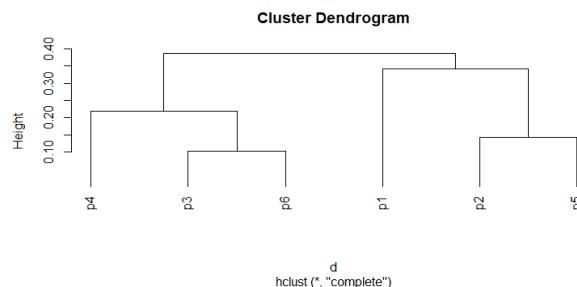


FIGURE 1.15 – Dendrogramme initial

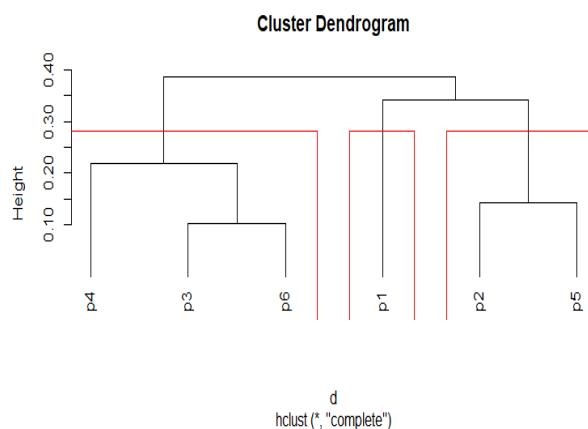


FIGURE 1.16 – Dendrogramme coupé

La coupure donne 3 classes $\{p_3, p_6, p_4\}$ et $\{p_2, p_5\}$ et $\{p_1\}$.

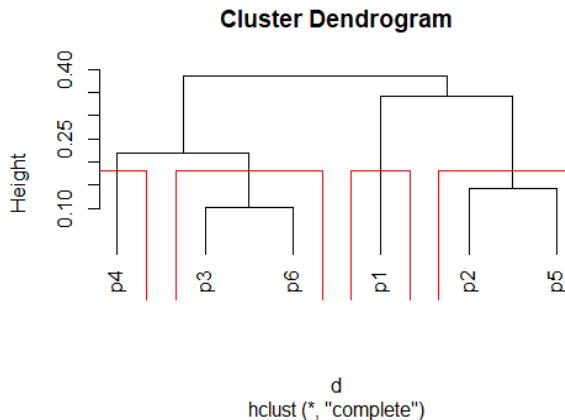


FIGURE 1.17 – Dendrogramme coupé

La coupure donne 4 classes $\{p_4\}$, $\{p_3, p_6\}$, $\{p_2, p_5\}$ et $\{p_1\}$.

Dans cette deuxième partie du paragraphe, on essaie d’interpréter les résultats obtenus.

Pour chaque classe C, on peut examiner :

- 1) Son effectif.
- 2) Son diamètre (distance entre les 2 points les plus éloignés), on le note, $\text{diam}(C)$.
- 3) La séparation (distance minimum entre la classe considérée C et la classe la plus proche) et le numéro de la classe la plus proche, on la note, $s(C)$.
- 4) Les identités des individus les plus proches du barycentre de la classe ou «parangons».
- 5) Les identités des individus les plus éloignés du barycentre de la classe ou «extrêmes».

En suite on peut faire une comparaison entre les différentes méthodes et voir les classes communes, les changements, d’une méthode à l’autre...etc

En regardant les dendrogrammes obtenus, dans l’exemple précédent, on constate qu’ils ont des formes différentes.

Dans le cas du critère du minimum, le dendrogramme a une forme en escaliers et les indices d’agrégation des données sont très proches entre elles sauf la donnée p_1 .

Dans le cas des méthodes utilisant le critère du maximum ou celui de la moyenne, la distribution des classes est bien distinguée.

Pour la méthode du maximum dans la FIGURE 1.16, on voit qu'il y a 3 classes.

- 1) La classe $C_1 = \{p_3, p_6, p_4\}$ de cardinal 3, la classe $C_2 = \{p_2, p_5\}$ de cardinal 2 et la classe $C_3 = \{p_1\}$ de cardinal 1.
- 2) $Diam(C_1) = d(p_4, p_6) = 0.22$, $Diam(C_2) = 0.14$ et $Diam(C_3) = 0$.
- 3) $s(C_2) = \min(d(C_2, C_1), d(C_2, C_3)) = d(C_2, C_3) = 0.15$, donc la plus proche classe de la classe C_2 est C_3 .
- 4) et 5) Parangons et extrêmes des classes :
 - Pour C_1 , on a un seul individu dans cette classe qui est aussi le barycentre : $g_1 = p_1$.
 - Pour C_2 , on a le barycentre de C_2 est le point $g_2(0.15, 0.395)$, $d(p_2, g_2) = 0.07$ et $d(p_5, g_2) = 0.07$.
 - Pour C_3 , on a le barycentre de C_3 est le point $g_3(0.35, 0.27)$, $d(p_3, g_3) = 0.05$, $d(p_4, g_3) = 0.12$ et $d(p_6, g_3) = 0.1$.

On peut aussi faire une comparaison des résultats obtenus par ces différentes méthodes. Pour le critère de la moyenne et celui du maximum, avec une coupure en 3 classes, on voit qu'ils gardent la même partition $C_1 = \{p_3, p_6, p_4\}$, $C_2 = \{p_2, p_5\}$ et $C_3 = \{p_1\}$, cependant le critère du minimum donne une partition différente, $C_1 = \{p_1\}$, $C_2 = \{p_2, p_3, p_5, p_6\}$ et $C_3 = \{p_4\}$.

Chapitre 2

Application numérique avec logiciel R

Dans ce chapitre on implémentera des différentes méthodes mentionnées dans le chapitre précédent, avec le logiciel R, sur un ensemble de données de fiches techniques des voitures, enregistré sous format ”.csv”, tiré du site web : <https://www.auto-selection.com/>.

Cet ensemble de données brutes est résumé dans les 4 tableaux suivants :

Sign in X

File Home Insert Page Layout Formulas Data Review View Help Tell me what you want to do

Share

L8 fx

A B C D E F G H I J K L M N O P Q

1	Marque	Nombre d	Longueur	Largeur en	Poids en	k	Nombre d	Vmax en	k	Accelerati	Consumm	Prix en euro
2	Bentley bentayga	5	5141	1998	2440	608	301	4.1	12.8	298990		
3	Porche 911 cabriolet	4	4507	1880	1670	580	330	3	9.3	221615		
4	Audi r8 spyder	2	4426	1940	1770	610	328	3.3	12.5	221600		
5	Ferrari california	4	4570	1910	1625	560	316	3.6	10.5	188353		
6	Nissan gtr	4	4690	1895	1800	600	315	2.8	11.8	184950		
7	Bentley continental gt	4	4818	1944	2295	507	305	4.8	10.6	182700		
8	Merecedes amg gt	2	4551	2007	1650	585	318	3.6	12.4	176400		
9	Audi r8	2	4426	1940	1640	540	320	3.5	11.4	173870		
10	Porche 911 coupe	2	4562	1852	1430	500	318	3.4	12.7	155255		
11	Porche 911 targa	4	4528	1852	1605	450	306	3.7	9.1	153215		
12	Audi rs7 sportback	4	5012	1911	1930	605	280	3.7	9.5	144550		
13	Audi q8	5	4986	1995	2145	286	245	6.3	6.8	85200		
14	Jeep grand cherokee	5	4821	1943	2403	250	202	8.2	7	74400		
15	Porche cayenne	5	4855	1939	2185	262	221	7.3	6.8	73610		
16	Maserati levante	5	5003	1968	2205	275	230	6.9	7.2	73200		
17	Audi q7	5	5052	1968	1995	218	216	7.1	5.5	71610		
18	Maserati ghibli	5	4971	1945	1875	275	250	6.3	5.9	68600		
19	Volvo xc90	5	4950	2008	1922	190	201	9.2	5.2	65849		
20	Vw touareg	5	4801	1940	2185	204	206	8.7	6.7	65730		
21	Land rover discovery	5	4970	2000	2105	180	189	10.5	6.2	64600		
22	Mercedes benz e break	5	4933	1852	1840	194	233	7.8	4.9	57650		
23	Mercedes benz e coupe	4	4826	1860	1805	194	239	7.6	5	57400		
24	Vw arteon	5	4862	1871	1753	240	245	6.5	5.9	55840		
25	Jaguar xf sportbrake	5	4955	1880	1660	163	219	9.3	4.5	53935		
26	Mercedes benz e	5	4923	1852	1680	194	240	7.3	3.9	49950		

Ready 100%

TABLE 2.1 – Données1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
26	Mercedes benz e	5	4923	1852	1680	194	240	7.3	3.9	49950							
27	Ford edge	5	4808	1928	1949	210	211	9.4	5.8	49800							
28	Jeep wrangle	4	4223	1873	1933	200	172	10.7	9	44500							
29	Jaguar xf	5	4964	1880	1545	163	229	8.7	4	41820							
30	Audi a6	5	4933	1874	1625	150	214	9.5	4.3	41460							
31	Vw amarok	5	5254	1954	2166	163	177	9.9	8.5	39468							
32	Ford kuga	5	4535	1856	1815	180	200	10	6	39100							
33	Bmw x2	5	4360	1824	1475	116	192	11.5	4.1	38850							
34	Opel insignia	5	4842	1856	1613	136	210	10.9	4.3	36260							
35	Bmw x1	5	4439	1821	1425	116	190	11.1	3.9	35400							
36	Jeep compas	5	3494	1819	1505	120	185	11	4.4	31800							
37	Hyundai tucson	5	4480	1850	1582	115	175	11.8	4.8	31400							
38	Ford focus	5	4382	1848	1493	150	210	8.5	4.4	31400							
39	Audi q2	5	4191	1794	1310	116	197	10.3	4.4	30180							
40	Vw tiguan	5	4486	1839	1499	115	185	10.9	4.7	30170							
41	Renault kadjar	5	4449	1836	1531	115	189	11.7	4.4	30100							
42	Honda civic	5	4370	1770	1425	120	201	10.8	3.4	29910							
43	Vw golf	5	4258	1790	1226	115	198	10.2	4.1	29740							
44	Nissan qashqai	5	4394	1806	1393	110	182	11.9	3.8	29700							
45	Skoda octavia break	5	4670	1814	1342	116	200	10.3	4	29690							
46	Mitsubishi asx	5	4365	1770	1380	114	182	11.2	4.6	28490							
47	Opel astra	5	4370	1809	1410	136	205	9.6	3.9	28450							
48	Mini cabrio	4	3821	1727	1265	116	195	9.9	4.2	28200							
49	Bmw serie1	5	4329	1765	1320	95	185	12.5	3.6	26850							
50	Audi a3	5	4241	1777	1290	116	202	10.2	4	26450							
51	Renault clio	5	4062	1732	1190	110	194	11.2	3.5	26400							

TABLE 2.2 – Données2

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
51	Renault clio	5	4062	1732	1190	110	194	11.2	3.5	26400							
52	Citroen berlingo van	5	4753	1921	1470	99	165	13	4.5	26220							
53	Mazada cx3	5	4275	1765	1200	105	177	10.1	4	25700							
54	Hyundai i30	5	4340	1795	1338	110	190	11	3.8	25650							
55	Alfa romeo guiliette	5	4354	1798	1310	120	195	10	3.8	25490							
56	Mitsubishi eclipse cross	5	4405	1805	1425	163	205	10.3	6.6	24990							
57	Peugeot 308	5	4253	1804	1160	100	186	11.3	3.6	24800							
58	Skoda octavia	5	4670	1814	1305	116	203	10.1	4.1	23490							
59	Fiat 500x	5	4248	1796	1320	95	172	12.9	4.1	23390							
60	Seat ibiza	5	4059	1780	1258	116	195	9.6	3.9	23045							
61	Renault megane	5	4359	1814	1205	90	175	13.4	3.7	22800							
62	Renault grand kangoo	5	4666	1829	1430	90	160	13.3	4.6	22750							
63	Jeep renegade	5	4255	1805	1404	120	178	10.2	4.4	22250							
64	Peugeot 2008	5	4159	1829	1205	102	182	10.6	4	21850							
65	Nissan juke	5	4135	1765	1347	110	175	11.2	4	21850							
66	Citroen c3 picasso	5	3996	1749	1090	99	185	11.9	3.7	21650							
67	Ford fiesta	5	4040	1735	1207	120	195	9	3.5	21650							
68	Dacia duster	5	3441	1804	1405	116	175	12.1	4.7	21350							
69	Fiat 500c	4	3571	1627	1020	95	180	10.7	3.4	21340							
70	Peugeot 208	5	3973	1829	1080	99	188	10.5	3.6	19600							
71	Hyundai i20	5	4035	1734	1165	75	159	16	3.5	19350							
72	Vw polo	5	3972	1682	1077	75	173	12.9	3.7	19230							
73	Toyota yaris	5	3945	1695	1085	90	175	10.6	3.5	18550							
74	Skoda fabia	5	3992	1732	1175	75	172	13.1	3.9	17790							
75	Opel corsa	5	4021	1746	1309	75	164	14.8	4	17750							
76	Fiat panda 4x4	4	3686	1672	1125	95	167	12.5	4.4	17490							

TABLE 2.3 – Données3

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
76	Fiat panda 4x4	4	3686	1672	1125	95	167	12.5	4.4	17490							
77	Honda jazz	5	3995	1694	1066	102	190	11.2	5	16710							
78	Dacia sandero	5	4089	1733	1083	90	166	12	3.8	14870							
79	Fiat doblo	5	4406	1832	1370	95	161	15.4	7.4	14490							
80	Dacia logan	5	4346	1733	1116	75	140	14.2	3.6	11710							
81	Kia picanto	5	3595	1595	953	67	161	14.3	4.4	11000							
82																	
83																	
84																	
85																	
86																	
87																	
88																	
89																	
90																	
91																	
92																	
93																	
94																	
95																	
96																	
97																	
98																	
99																	
100																	
101																	

TABLE 2.4 – Données4

Notre objectif, dans cette partie est de faire une segmentation sur cet ensemble de voitures de sorte à obtenir des catégories naturelles.

Dans un premier temps, on importe les données dans RStudio (environnement de travail) avec la commande : `"dt <- read.csv(file.choose())"` et on commence les étapes de prétraitement des données (le centrage et la réduction) des variables quantitatives avec la commande : `"dt1 <- scale(dt[, -1])"`. On n'a pas de données manquantes à traiter, dans notre ensemble de données.

Avant tout, on essaie de décrire et de visualiser les données brutes (i.e à l'état initial) avec la commande `"pairs(dt)"`.

L'ensemble des données comporte 80 voitures de différents marques et modèles et 9 variables descriptives : Nombre de places, longueur en mm, largeur en mm, poids en kg, nombre de chevaux, vitesse maximale (Vmax) en km/h, accélération de 0-100 km/h/seconde, consommation en litres/100 km, prix en euro.

On visualise l'ensemble des données à travers la figure suivante :

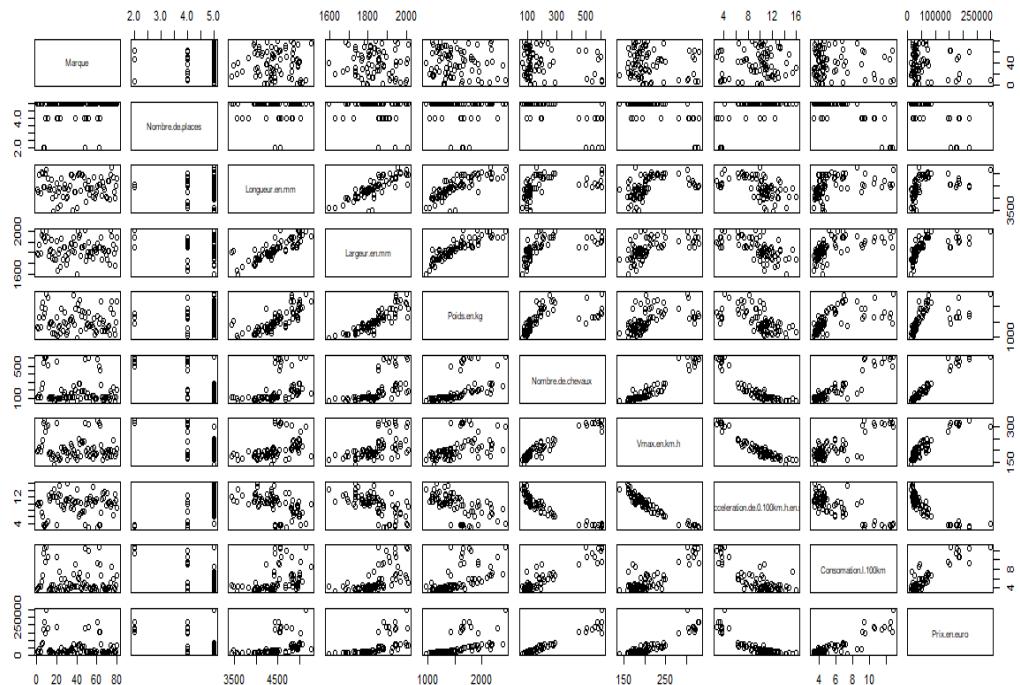


FIGURE 2.1 – Visualisation des données deux à deux

Dans la FIGURE 2.1, on voit les différentes répartitions des données en tenant compte de deux variables à la fois. En considérant, par exemple :

- 1) Le rectangle situé dans l'intersection de la ligne 4 avec la colonne 5, on constate qu'il y a un seul groupe contenant tous les individus en examinant le graphe de la largeur en fonction du poids.
- 2) Le rectangle situé dans l'intersection de la ligne 7 avec la colonne 8, on remarque qu'il y a deux groupes en examinant le graphe de la vitesse maximale en fonction de l'accélération.
- 3) Le rectangle situé dans l'intersection de la ligne 4 avec la colonne 2, on voit qu'il y a trois groupes en examinant le graphe de la largeur en fonction de nombre de places.
- 4) Le rectangle situé dans l'intersection de la ligne 8 avec la colonne 5, on ne peut rien dire en examinant le graphe de l'accélération en fonction du poids. La description diffère d'un rectangle à l'autre dans la FIGURE 2.1. Une question qui se pose : Combien de groupes peut-on considérer ?

Pour répondre à cette question on fait une classification hiérarchique ascendante avec différentes méthodes et on interprète les résultats obtenus.

a) Méthode de Ward :

En suivant les étapes de l'algorithme de la classification hiérarchique ascendante, on commence par le calcul de la matrice des distances entre les individus, avec la commande : " $d < -dist(dt1)$ ". On fait intervenir la fonction "hclust" qui implémente l'algorithme choisi.

On exécute donc la commande suivante : " $h.w < -hclust(d, "ward.D2")$ ", puis on visualise les résultats de sortie de la fonction "hclust" avec la commande : " $plot(h.w, labels = dt[,1], hang = -1, cex = .6)$ ". On obtient la figure suivante :

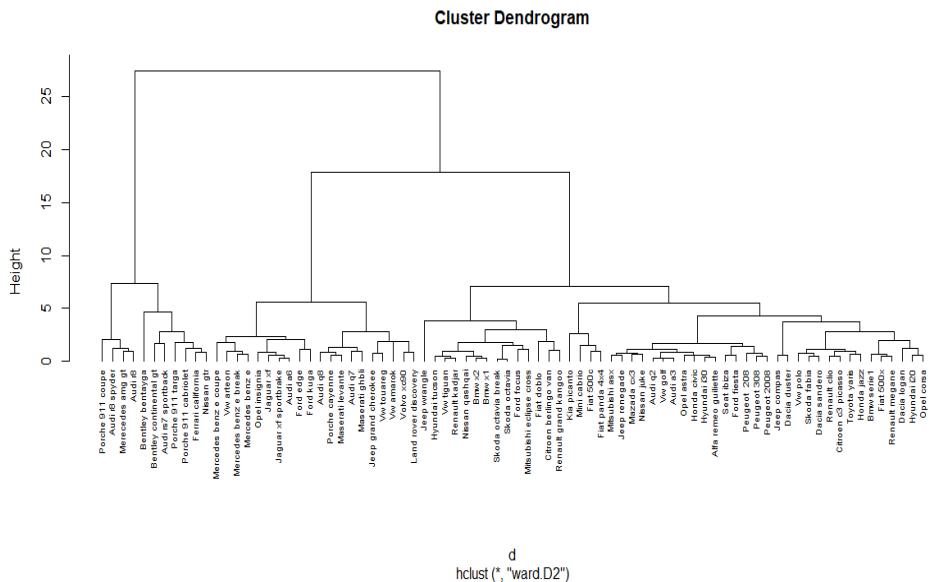


FIGURE 2.2 – Dendrogramme relatif au critère de Ward

Un premier coup d'oeil sur la FIGURE 2.2, nous montre que les marques sont réparties sur l'axe des abscisses et forment différents groupes. Pour bien visualiser la répartition on exécute la commande : " $rect.hclust(h.w, 3)$ ".

Cette commande nous donne une coupure en 3 groupes, comme présenté dans la figure suivante :

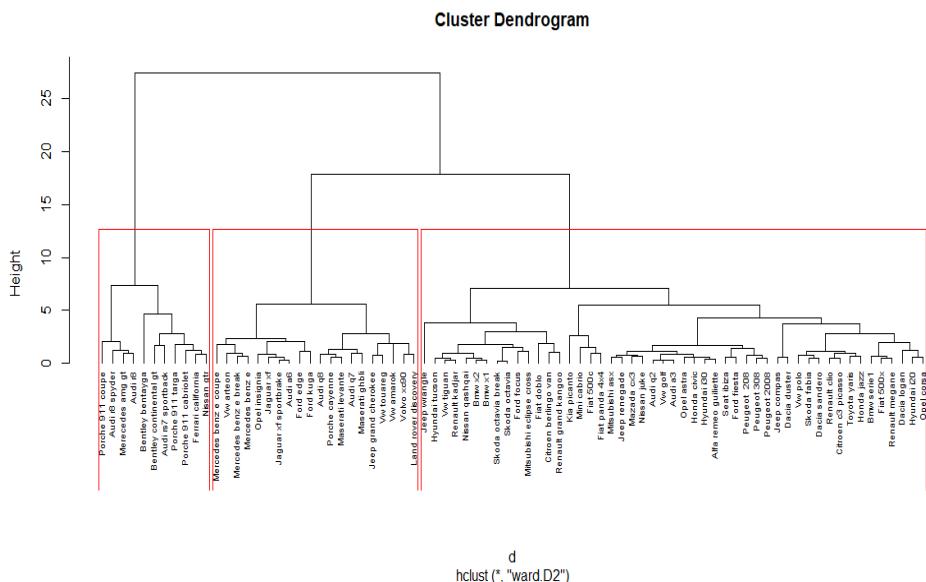


FIGURE 2.3 – Dendrogramme coupé relativement au critère de Ward

Dans la FIGURE 2.3, on voit 3 classes, la première classe contient les voitures dites, voitures de hautes performances, la deuxième classe contient des voitures de milieu de gamme avec des performances avancées et la troisième classe contient les voitures d'entrée de gamme avec des performances économiques.

Si on coupe la hiérarchie un peu plus haut, on trouvera deux classes, une contenant les voitures de luxe et l'autre contenant les voitures économiques. Si maintenant, on coupe un peu plus bas, on trouvera beaucoup plus de classes que dans les cas précédents et donc on trouvera de nouvelles catégories.

Le choix du nombre de classes est une tâche délicate. Des fois, la réponse est clairement visible sur le dendrogramme, mais dans notre situation, on peut considérer 2 classes comme on peut en considérer 3 ou bien 4 ...etc. Pour préciser le choix, on utilise la méthode du coude qu'on implémente, on obtient ainsi la figure suivante :

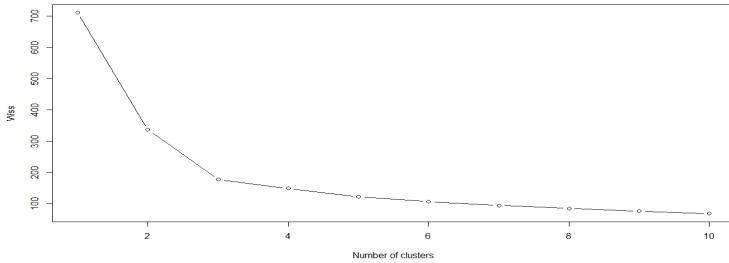


FIGURE 2.4 – Implémentation de la méthode du coude

On voit dans la FIGURE 2.4 que le point de déviation aigüe donne 3 classes à considérer (i.e la coupure définie dans la FIGURE 2.3), la première classe C_1 des voitures de hautes performances, située à gauche de la FIGURE 2.3, la deuxième classe C_2 contient plus de voitures de milieu de gamme, située au milieu de la FIGURE 2.3 et une troisième C_3 contient plus de voitures d'entrée de gamme, située à droite de la FIGURE 2.7. On voit que la classification ici n'est pas totalement exacte, mais apparemment, elle est meilleure que la classification manuelle. La qualité de la partition en 3 classes est donnée par l'indicateur $R^2 = 75\%$, qui est beaucoup mieux.

On tente d'interpréter les résultats obtenus pour mieux comprendre la répartition des données selon cette méthode.

1) On considère une partition en 3 classes, on exécute les commandes : "clusters.w <- cutree(h.w, 3)" et "clusters.w", on trouve le regroupement suivant :

```
> clusters.w
[1] 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 3 3 2 3 3 3 3
[38] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[75] 3 3 3 3 3
>
```

TABLE 2.5 – Description des classes

Comme on le voit ci-dessus dans le TABLE 2.5, on constate un chevauchement dans la première ligne entre les voitures de la deuxième classe et celles

de la troisième classe. Pour mieux voir les classes, on exécute la méthode de silhouette pour examiner la pertinence des individus à leurs classes, avec les commandes : "silhouette(clusters.w,d)" et "plot(silhouette(clusters.w,d))". On obtient ainsi la figure suivante :

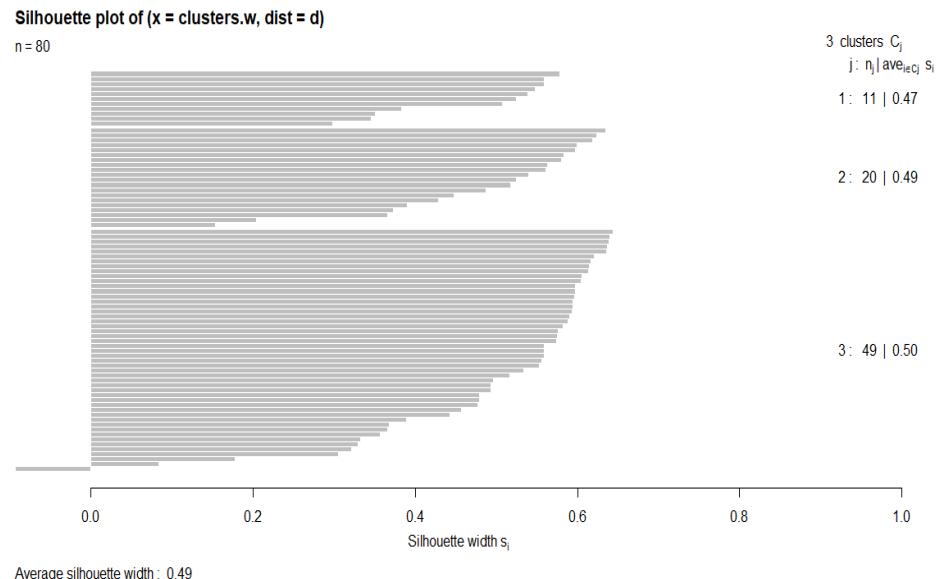


FIGURE 2.5 – Description des classes selon les indices de silhouette

Dans la FIGURE 2.5, la méthode de silhouette est implémentée pour une coupure en 3 classes, C_1 , C_2 et C_3 , les descriptions des 3 classes apparaissent sur la même FIGURE 2.5.

L'examen de la pertinence des individus à la classe C_1 est réalisé à l'aide des commandes :"sil.w <- silhouette(clusters.w,d)" et "sil.w". Le résultat obtenu est le suivant :

	cluster	neighbor	sil_width
[1,]	1	2	0.29774005
[2,]	1	2	0.52479481
[3,]	1	2	0.57850426
[4,]	1	2	0.53894261
[5,]	1	2	0.55855822
[6,]	1	2	0.38350615
[7,]	1	2	0.54823010
[8,]	1	2	0.55901771
[9,]	1	2	0.50766007
[10,]	1	2	0.35064705
[11,]	1	2	0.34543937

TABLE 2.6 – Pertinence des individus de la classe C_1

D'après le TABLE 2.6, les indices de silhouettes des individus de la classe C_1 indiquent que tout les points de cette classe ont pour classe voisine la classe C_2 . La classe C_1 se trouve bien isolé dans cette partition en trois classes, car tous les indices sont strictement positifs et même supérieurs à 0.2. Ceci est raisonnable car cette classe contient des voitures de hautes performances et a un prix très élevé par rapport aux autres classes.

De même pour la classe C_2 , on a le tableau des pertinences suivant :

	cluster	neighbor	sil_width
[1,]	2	3	0.5577431
[2,]	2	3	0.6238268
[3,]	2	3	0.6196222
[4,]	2	3	0.6408295
[5,]	2	3	0.6027882
[6,]	2	3	0.5933702
[7,]	2	3	0.5658530
[8,]	2	3	0.5982595
[9,]	2	3	0.5343908
[10,]	2	3	0.5422696
[11,]	2	3	0.5962887
[12,]	2	3	0.4112619
[13,]	2	3	0.5288645
[14,]	2	3	0.4628476
[15,]	2	3	0.4507124
[16,]	2	3	0.5837909
[17,]	2	3	0.3923674
[18,]	2	3	0.3820394
[19,]	2	3	0.4886323
[20,]	2	3	0.2081525
[21,]	2	3	0.1632428
>			

FIGURE 2.6 – Pertinence des individus de la classe C_2

Le TABLE 2.6 indique que cette classe a pour voisine de tous ses points, la classe C_3 et elle est bien isolée, car les indices sont strictement positifs et même supérieurs à 0.1.

Pour la classe C_3 , on obtient les deux tableaux des pertinences suivants :

	cluster	neighbor	sil_width
[1,]	3	2	-0.09296661
[2,]	3	2	0.47904301
[3,]	3	2	0.47880365
[4,]	3	2	0.45710329
[5,]	3	2	0.30485823
[6,]	3	2	0.17748160
[7,]	3	2	0.59350350
[8,]	3	2	0.35676243
[9,]	3	2	0.38901300
[10,]	3	2	0.53374585
[11,]	3	2	0.59611894
[12,]	3	2	0.55852274
[13,]	3	2	0.32976348
[14,]	3	2	0.58242653
[15,]	3	2	0.44282753
[16,]	3	2	0.49640766
[17,]	3	2	0.61470281
[18,]	3	2	0.59414574
[19,]	3	2	0.63541278
[20,]	3	2	0.08336389
[21,]	3	2	0.61632069
[22,]	3	2	0.59435086
[23,]	3	2	0.55297301
[24,]	3	2	0.32185779
[25,]	3	2	0.62076865
[26,]	3	2	0.33280426
[27,]	3	2	0.61343866
[28,]	3	2	0.59709577
[29,]	3	2	0.57393505

FIGURE 2.7 – Pertinence des individus de la classe C_3

[29,]	3	2	0.57393505
[30,]	3	2	0.36809053
[31,]	3	2	0.55871981
[32,]	3	2	0.59062929
[33,]	3	2	0.63935397
[34,]	3	2	0.64424586
[35,]	3	2	0.57569382
[36,]	3	2	0.49332728
[37,]	3	2	0.47739379
[38,]	3	2	0.57500585
[39,]	3	2	0.55859376
[40,]	3	2	0.60458952
[41,]	3	2	0.60540015
[42,]	3	2	0.63636092
[43,]	3	2	0.58857020
[44,]	3	2	0.51622659
[45,]	3	2	0.59700019
[46,]	3	2	0.63964774
[47,]	3	2	0.36546429
[48,]	3	2	0.55602100
[49,]	3	2	0.49352683

FIGURE 2.8 – Suite du "TABLE 2.7"

D'après le TABLE 2.7 et le TABLE 2.8, on voit que tout les points de la classe C_3 ont pour classe voisine la classe C_2 . Il y a aussi un individu d'indice de silhouette négatif, ce qui signifie que cet individu n'est pas cohérent avec la classe C_3 mais avec la classe C_2 , il s'agit de la voiture "Jeep wrangle" qui a des performances avancées, donc il est clair que cette voiture est mal classé, malgré ceci, si on compte tenu la moyenne des indices lors de chaque

classes, $S(C_1) = 0.43$, $S(C_3) = 0.50$ et $S(C_3) = 0.50$, donc C_2 et C_3 sont plus homogènes que C_1 . Ceci revient à ce que les voitures dans C_1 sont un peu distinctes et ceci est montré dans la FIGURE 2.5 avec les sauts importants des indices de silhouettes des individus, ce qui est confirmé par ses attributs qui sont relativement distincts, par contre pour la classe C_2 et C_3 , on voit qu'il y a une augmentation souple des indices de silhouette des individus dans la même FIGURE 2.5 et ceci revient aux attributs relativement communs entre ses voitures.

On peut aussi implémenter une méthode basée sur l'indice de silhouette pour déterminer le nombre de classes à considérer, on a ainsi la figure suivante :

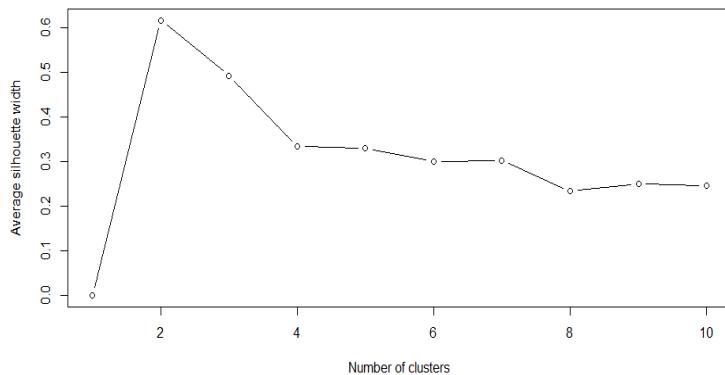


FIGURE 2.9 – Méthode de Silhouette

D'après la FIGURE 2.9, on constate que la méthode indique 2 classes à considérer, mais comme cela a été mentionné auparavant, on cherche toujours des partitions autres que la partition en 2 classes, qui est dans la plupart des cas, une partition évidente.

Dans le but d'avoir une bonne classification, on examine la variation des partitions d'un critère à l'autre, ce qu'on appelle "Crossed Validation" et on regarde le critère le plus adapté à ce type de données.

b) Méthode du maximum :

Le résultats correspondent à l'utilisation de l'algorithme de cette méthode donne le dendrogramme suivant :

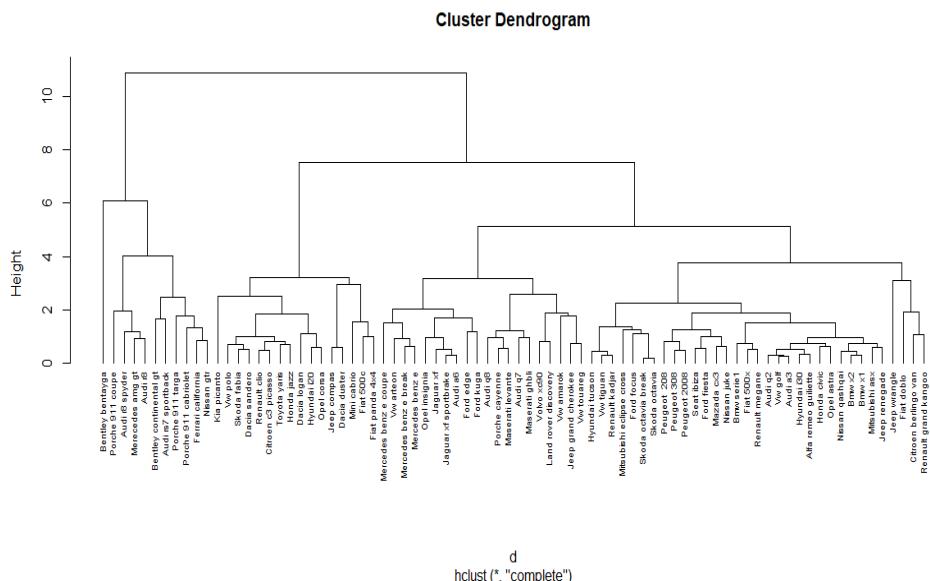


FIGURE 2.10 – Dendrogramme relatif à la méthode du maximum

Dans cette FIGURE 2.10, on voit une répartition des individus différente de celle de la méthode de Ward. On exécute une coupure en 3 classes. On obtient :

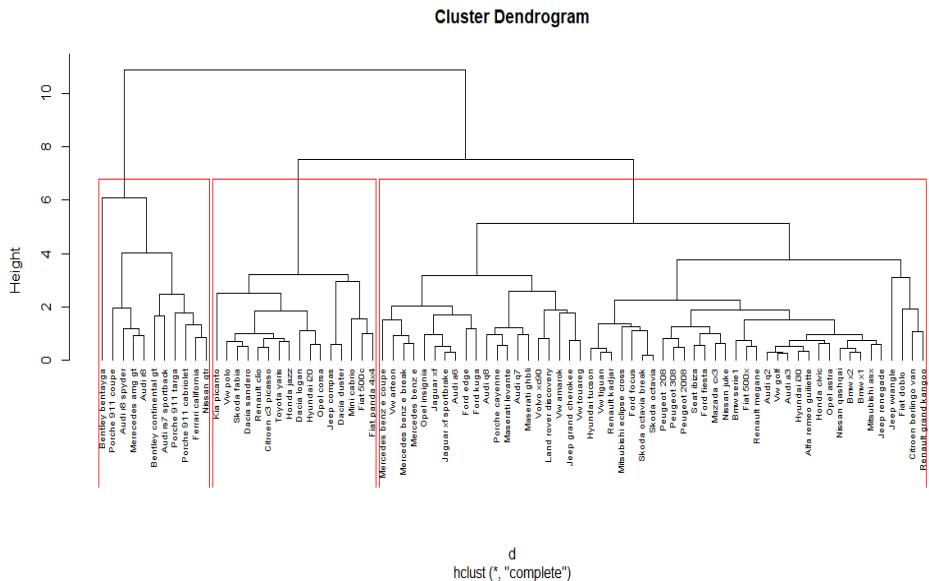


FIGURE 2.11 – Dendrogramme coupé relatif à la méthode du maximum

Dans la FIGURE 2.11, on voit que cette méthode a réussi à mieux classifier la première classe comme l'a fait la méthode de Ward, mais elle donne un chevauchement de la deuxième et de la troisième classe. Ceci vient du fait que cette méthode est biaisée vers les petites classes. La première classe se trouve assez loin des deux autres classes et ne peut donc se chevaucher avec elles. Ceci est confirmé par la méthode de silhouette, qui montre la pertinence des individus dans la figure suivante :

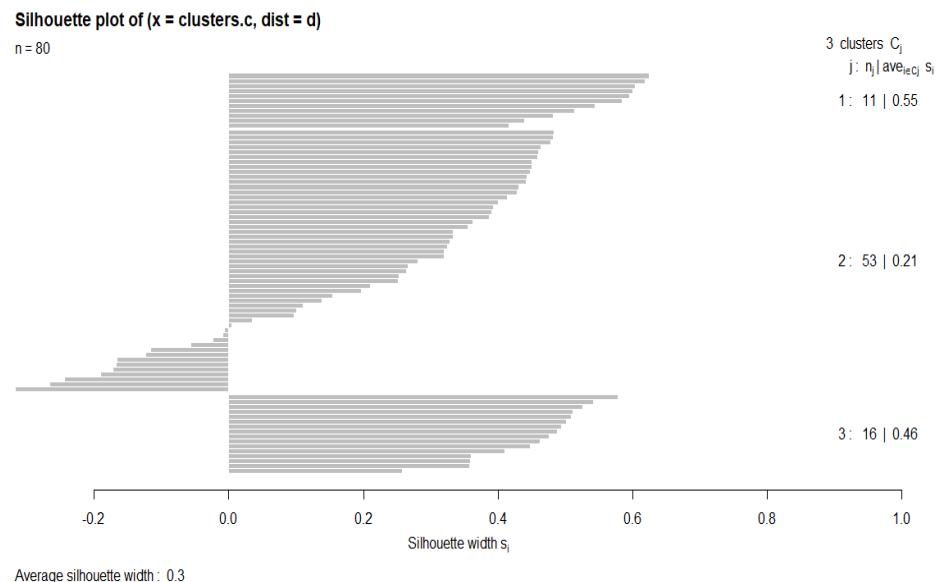


FIGURE 2.12 – Description des classes selon les indices de silhouette

Comme l'indique la FIGURE 2.12, on voit clairement le positionnement négatif ou presque nul d'une portion importante dans la deuxième classe. D'après les informations contenues dans les indices de silhouette, on voit que la classe voisine des individus d'indice négatif est la classe 3, ce qui confirme notre interprétation ci-dessus.

On utilise maintenant la méthode du coude pour avoir le nombre de classe à considérer. On obtient le graphe suivant :

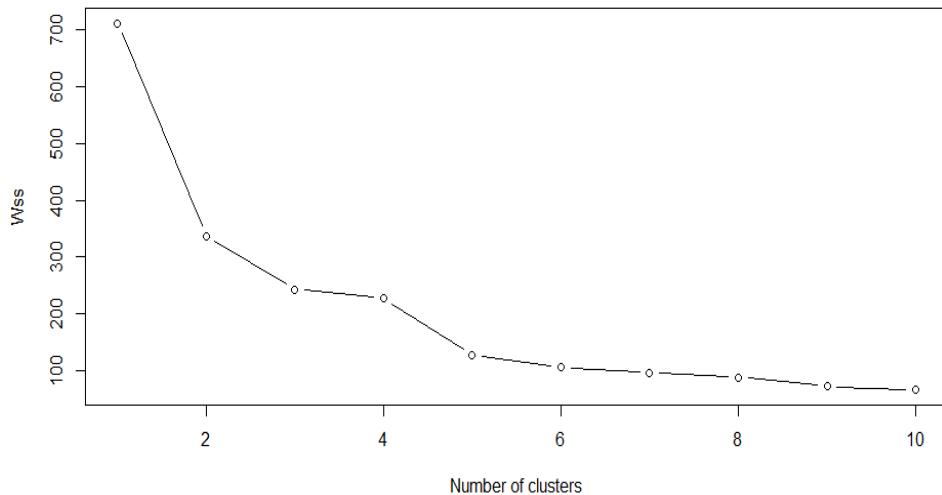


FIGURE 2.13 – Méthode du coude

D'après la FIGURE 2.13, on voit que la courbe présente plusieurs déviations, une augue, une douce puis une plus importante. On peut dire que la méthode a réussi de classifier les classes naturellement dans la partition en 5 classes. Cette partition en 5 classes est presque la même que la partition obtenue par la méthode de Ward en 5 classes.

c) Méthode du minimum :

Comme cela a été fait pour les méthodes précédentes, on examine les résultats obtenus par cette méthode, on commence par donner le dendrogramme correspondant aux données :

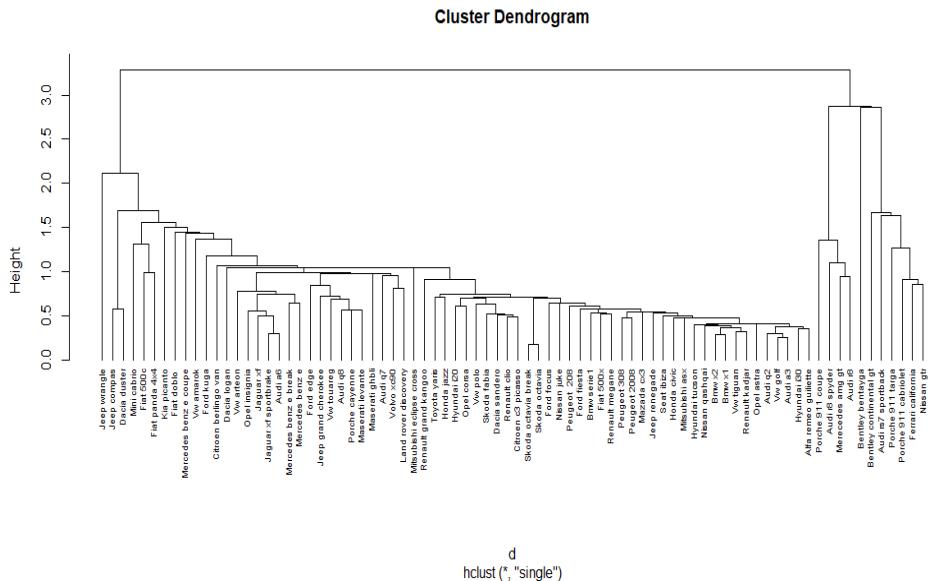


FIGURE 2.14 – Dendrogramme relatif à la méthode du minimum

La FIGURE 2.14, montre essentiellement deux classes, l'une est grande et sous forme d'escaliers et l'autre est plus petite. Pour plus de précision, on effectue la coupure suivante, du dendrogramme ci-dessus et on obtient :

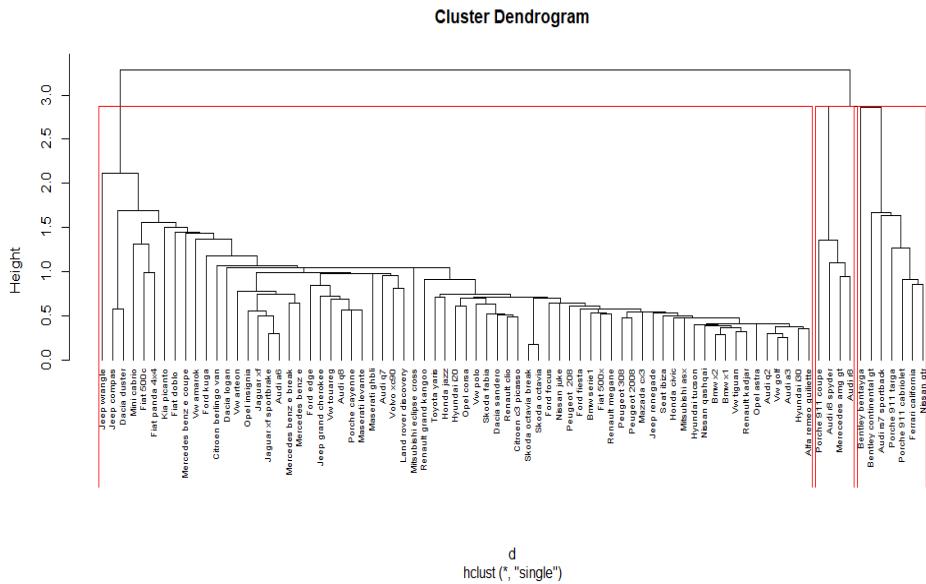


FIGURE 2.15 – Dendrogramme coupé relatif à la méthode du minimum

Ici la situation diffère de la méthode du maximum. D’après la FIGURE 2.15, on constate que la coupure donne une petite classe coupée en deux sous-classes et une grosse classe. Cette méthode étant sensible à l’effet de chaîne, elle donne un chevauchement de la classe naturelle 2 des voitures de milieu de gamme avec la classe naturelle 3 des voitures d’entrée de gamme et elle partage la classe naturelle 1 en deux.

En utilisant, l’indice de silhouette, on obtient le diagramme suivant :

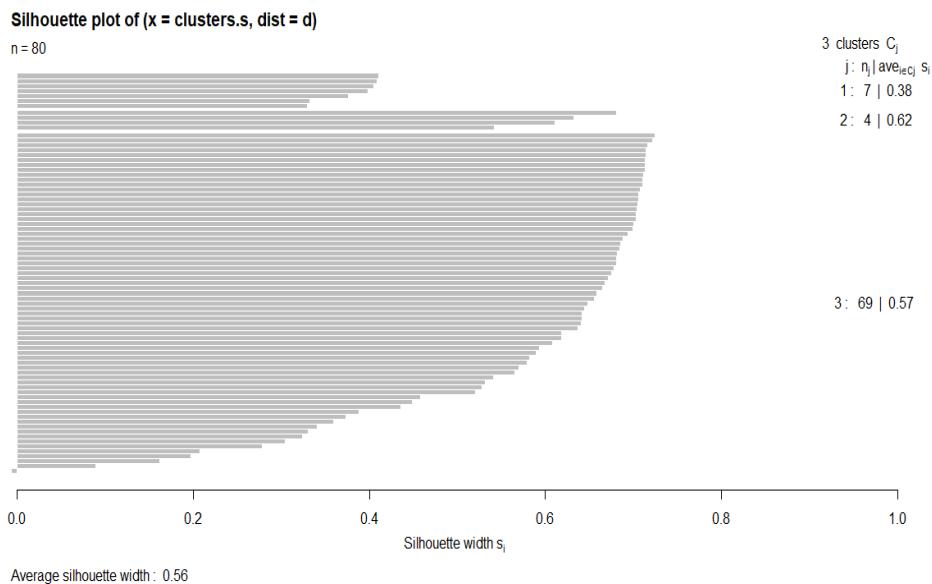


FIGURE 2.16 – Description des classes selon les indices de silhouette

On voit sur la FIGURE 2.16 que la grosse classe est très homogène. Ceci vient du fait que la méthode a fait une classification en deux classes et après elle a coupé la première classes en deux.

En utilisant la méthode du coude, on obtient :

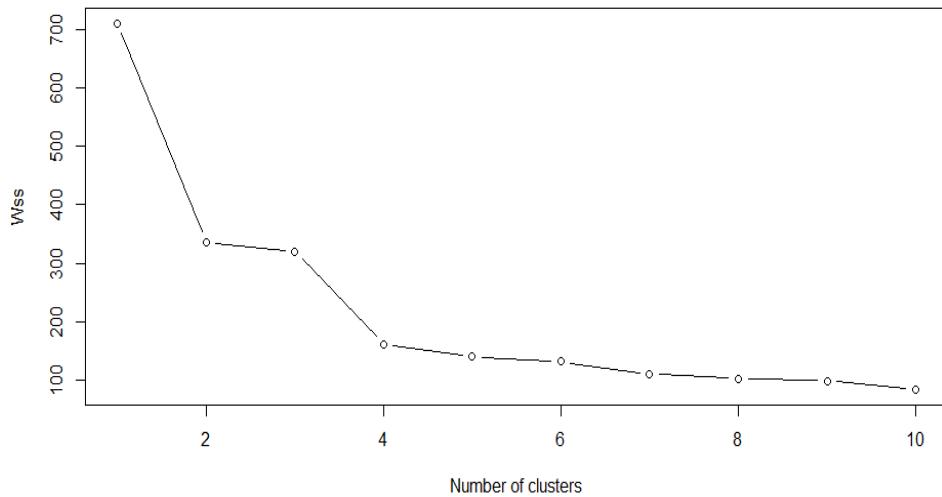


FIGURE 2.17 – Description des classes selon les indices de silhouette

On voit sur la FIGURE 2.17, une déviation de la courbe suivie d'une presque stabilisation et juste après il y a une autre déviation, pour les 4 premières coupures, on voit sur le dendrogramme que la grosse classe n'est pas touché, ce qui signifie que les individus de cette classe sont fortement emboîtés.

Interprétation générale :

En examinant les trois méthodes ci-dessus, on peut constater qu'elles ont des points communs et d'autres différents.

- 1) Points communs : Les trois méthodes ont réussi à isoler la classe des voitures de hautes performances, elles indiquent que les individus de cette classe sont un peu distincts entre eux, mais cette classe est loin d'autres classes.
- 2) Points de distinction : Les trois méthodes classifient la deuxième et la troisième classes de manières différentes. La méthode de Ward a séparé ces deux classes, en une classe de milieu de gamme avec des individus ayant des performances avancées et en une autre classe d'entrée de gamme avec des caractéristiques économiques. Seule une voiture de milieu de gamme est mal

classée (Jeep wrangle), ses dimensions ont joué un rôle dans sa mal classification, car cette voiture possède de petites dimensions. La méthode du maximum, a elle aussi, séparé ces deux classes mais d'une manière biaisée vers la classe de milieu de gamme. La méthode du minimum a considéré ces deux classes comme étant une seule classe. Les individus de ces deux classes sont fortement emboîés.

conclusions :

1) Il est connu que l'ensemble des voitures se répartit généralement en 3 classes, dans notre cas on a :

a) la classe des voitures de haute de gamme :

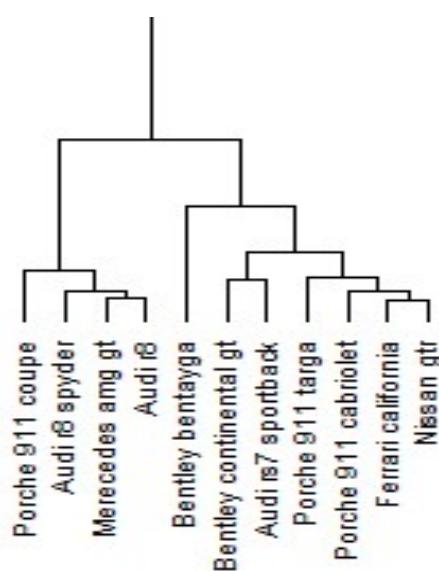


FIGURE 2.18 – Classe des voitures de haute de gamme

Cette classe dans est confirmée par les trois méthodes.

b) La classe de milieu de gamme :

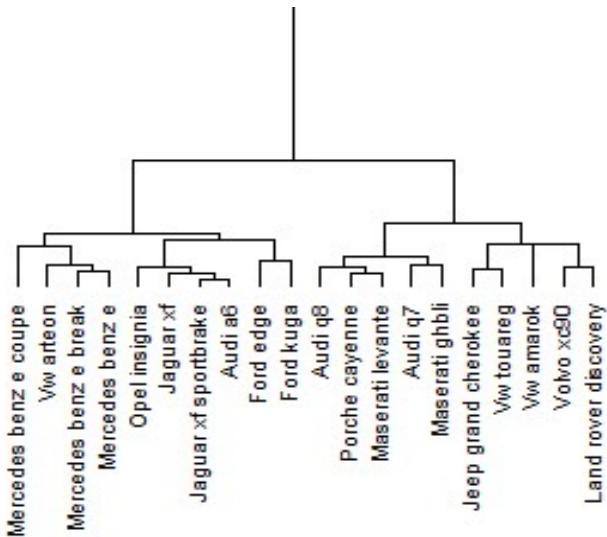


FIGURE 2.19 – Classe des voitures de milieu de gamme

Cette classe est confirmée par la méthode de Ward, on trouve aussi cette classe dans la partition en 5 classes dans la méthode du maximum. En se basant sur l'indice de silhouette dans la méthode de Ward, on ajoute la voiture Jeep wrangle à cette classe.

c) La classe d'entrée de gamme :

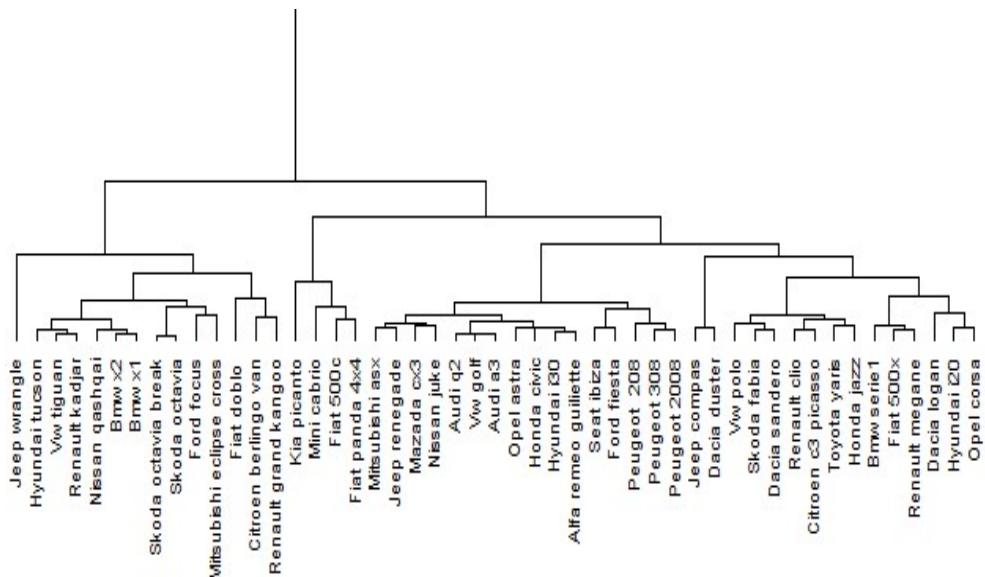


FIGURE 2.20 – Classe des voitures d'entrée de gamme

Cette classe est confirmée par la méthode de Ward, en enlevant la voiture Jeep wrange et par la méthode du maximum en tenant compte les individus d'indices de silhouette négatifs.

La méthode du minimum nous a donnée des information sur la distinction des individus dans la première classe et sur l'emboîtement des individus des deux autres classes.

2) La méthode de Ward paraît plus performante dans cette classification que les autres méthodes, car elle a pu faire une classification en 3 classes, qui est la classification naturelle avec une pertinence des individus cohérents.

Conclusion

La classification hiérarchique ascendante est une méthode intéressante. Elle permet de déterminer les liens naturels entre les individus, d'une manière automatique et elle permet de visualiser clairement ces liens par des dendrogrammes. De plus elle possède plusieurs méthodes adéquates pour différentes situations. On peut même utiliser la "Crossed Validation" (comparaison entre les méthodes). Cependant, l'inconvénient de cet algorithme est le temps d'exécution qui est en générale de l'ordre de $(O(n^3))$ et l'espace mémoire nécessaire qui est de l'ordre de $(O(n^2))$. Cet ordre de complexité rend le traitement des ensembles de données de grandes tailles assez lourd, ce qui ouvre la porte à l'implémentation d'autres algorithmes de complexité moins élevée tel que, l'algorithme des "K-means" ou bien l'implémentation mixte des deux algorithmes, l'algorithme des "K-means" pour le démarage de la classification suivi de l'algorithme de la classification hiérarchique ascendante.

Bibliographie

- [1] E. Lebarbier, T. Mary-Huard, *Classification non supervisée*, AgroParis-Tech.
- [2] François Husson, *Classification ascendante hiérarchique (CAH)*, Laboratoire de mathématiques appliquées - Agrocampus Rennes.
- [3] Gabor J.Szekely and Maria L. Rizzo, *Hierarchical clustering via Joint Between-Within Distances : Extending Ward's Minimum Variance Method*, Journal of Classification, vol. 22, no 2, septembre 2005, p. 151-183.
- [4] Chaitanya Reddy Patlolla, *Understanding the concept of Hierarchical clustering Technique* , <https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec>.
- [5] Peter J. ROUSSEEUW, *Silhouettes : a graphical aid to the interpretation and validation of cluster analysis* , Journal of Computational and Applied Mathematics 20 (1987) 53-65 North-Holland, University of Fribourg, ISES, CH-I 700 Fribourg Switzerland.