République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme **:**

MASTER en Mathématiques

Option : **Statistique**

Par

BENZIADI Chahla

Titre :

Méthodes de Classification

Membres du Comité d'Examen :

Dr.	**BENELMIR Imane**	UMKB	Encadreur
Dr.	**ABDELLI Jihane**	UMKB	Président
Dr.	**DHIABI Samra**	UMKB	Examinateur

Juin 2021

Dédicace

Je dédie ce mémoire

À mes chers parents,

À mon frère,

À ma soeur

À mon encadreur

À toutes mes amies

À tous ceux qui m'ont soutenu et aidé.

REMERCIEMENTS

Au nom du Dieu le plus clément et le plus miséricordieux.
Avant toute chose, je remercie **ALLAH** le tout le puissant de m'avoir donné la force pour l'achèvement de ce travail et de m'avoir donné le courage et la patience durant toutes ces années d'études.

Je tiens à exprimer ma gratitude à mon encadreur **D.r**. **Benelmir Imane** d'avoir accepter diriger de ce travail, pour son précieux soutien, ses remarques et ses conseils constructifs, sa patience et ses connaissances. Sans tout cela, ce mémoire n'aurait pas pu être réalisé. La confiance qu'elle m'a témoignée, sa grande disponibilité et ses judicieux conseils ont assuré la réussite de ce projet. Un remerciement spécial aussi pour avoir lu attentivement mon mémoire et avoir suggéré des modifications ce qui m'a permis d'améliorer sa qualité.

Mes vifs remerciements vont également aux membres du jury qui m'ont fait l'honneur d'examiner et d'évaluer mon travail avec le poids de leurs compétences.

Je remercie ma mère et mon père **Aziza** et **Bouzid** pour tous leurs sacrifices, leurs amours, leurs tendresses, leurs soutiens, leurs prières tout le long de ma vie.

Je remercie mon cher frère **Mohamed** pour son suivi continu tout au long des années d'études.

Je remercie ma chère soeur **Nadjet** pour ses encouragement permanent et son soutien moral.

Je remercie toutes mes amies **Sara, Kholoud, Rachida, Radhia, Khaoula**, **Asma** et **Mina.** Merci d'être une partie merveilleuse de ma vie.

Table des matières

Table des figures

Liste des tableaux

Introduction

La classification n'est pas un domaine d'application nouveau. En fait, elle est l'une des tâches les plus anciennes de la recherche d'information. Ses débuts remontent aux années 1960.

Dans mes recherches, je concentre essentiellement sur la classification non supervisée ou clustering qui s'agit d'extraire des classes ou groupes d'individus présentant des caractéristiques communes sans aucune connaissance préalable. Dans cette mémoire j'ai tenté de répondre à la problématique suivante :

Comment procèdé à la méthode de classification ?

Cette problématique renvoie à un certain nombre d'interrogation :

– Quelles sont les méthodes de classification ?

– Comment on représente les données ?

– Quelles méthodes de classification doit-on utiliser sur ces données ?

– C'est quoi le critère de classification ?

La classification est une des méthodes statistiques largement utilisées, elle joue un rôle important dans divers domaines comme en sciences biologiques (zoologie et écologie), médecine (survie, décès), chimie (tableau périodique des éléments), ect.

Le but de la classification est de regrouper des objets possédant des propriétés similaires en plusieurs classe homogènes et à séparer celles qui sont dissimilaires comme le montre la figure1.

Ce mémoire ces compose en deux chapitres :

Dans le premier chapitre on va discuter sur le principe de la méthode de classification hiérarchique on passant d'abord par un rappel sur des notions nécessaires, de finir un tableau des données, des mesures d'éloignement , puis décrire les deux types de classification hiérarchique à la savoir la classification ascendante hiérarchique (CAH) et la classification descendante hiérarchique (CDH) en implémentant ces deux dernières par leurs algorithmes.

Le deuxième chapitre est devisé en deux parties le première partie est sur la classification non hiérarchique on va expliquer le principe de cette méthode et présenter ces différents algorithmes qui sont : algorithme de K-means et algorithme des centres mobiles et enfin algorithme des nuées dynamiques.

Tondis que la deuxième partie est consacré à l'application qui sera exécutée sur logiciel R.

On clôture ce modeste travail par un conclusion générale.

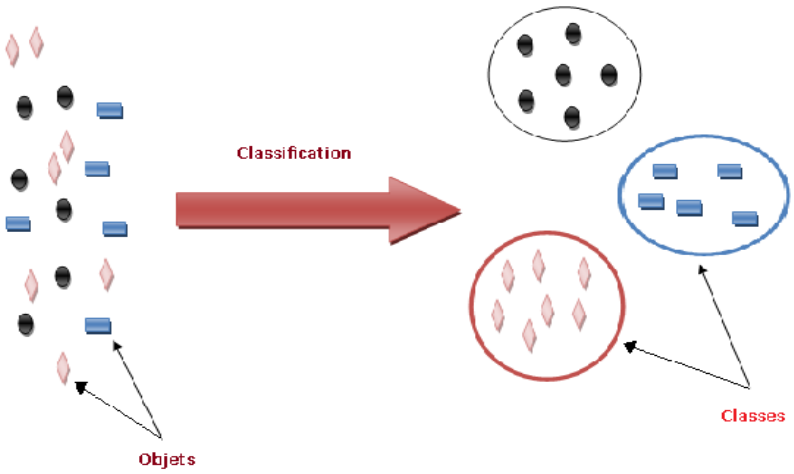

FIG. 1 – Classification.

Chapitre 1

Classification hiérarchique

La classification hiérarchique est a priori est depuis longtemps une problématique importante, elle est utilisée pour étudier des phénomènes naturelles comme la biologie en particulier. Le terme anglais pour la classification hiérarchique est "hierarchical clustering". L'objectif de cette méthode est la constitution des groupes d'individus aussi similaires que possibles, elle a l'avantage d'être interprétable à l'aide des arbres hiérarchique. Il existe deux types de méthodes hiérarchiques : la méthode ascendante et la méthode descendante.

- L'approche ascendante démarre avec chaque objet formant une classe distincte. On fusionne à chaque étape les deux classes les plus proches afin de n'obtenir qu'une seule classe.
- L'approche descendante démarre avec tous les objets dans une seule et même classe. A chaque itération, une classe est décomposée en plus petites classes, jusqu'à n'avoir plus qu'un seul objet dans chaque classe.

1.1 Principe de la classification hiérarchique

Le principe de la classification hiérarchique est de grouper des éléments proches dans un même groupe soient le plus similaires possible, c'est aussi la création d'une série de divisions pour un groupe d'individus, selon un standard prédéterminé de similitude exprimé sous la forme d'une matrice de distance, exprimant la distance entre chaque individu deux à deux.

On dira que des individus se ressemblent si les points associés sont proches les uns des autres (si les distances qui les séparent sont petites). La classification hiérarchique regroupe ou divise à plusieurs reprises les individus pour produire l'arbre de classification.

1.2 Données et leurs caractéristiques

Les données se représentent généralement sous la forme d'un tableau rectangulaire, dont les lignes correspondent à des individus ou unités statistiques et les colonnes à des variables appelées caractères ou caractéristique.

1.2.1 Tableau des données

L'ensemble des valeurs x_{ij} est présenté sous la forme d'une matrice X, de n lignes et k colonnes, appelée tableau des données.

$$X := (x_{ij})_{n\times k} = \begin{bmatrix} x_{11} & x_{12} & & x_{1k} \\ x_{21} & x_{22} & \ddots & x_{2k} \\ & & & \\ x_{n1} & x_{n2} & & x_{nk} \end{bmatrix} \in M_{n,k}(\mathbb{R}).$$

Remarque 1.2.1 *$\forall i = 1, \ldots, n$ et $j = 1, \ldots, k$, on a.*

1. Chaque variable peut être représentée par un vecteur de dimension n, appelé *vecteur variable*, correspondant aux valeurs prises par cette variable sur les n individus. On la note par

$$\boxed{x_j = (x_{1j}, \ldots, x_{nj})^t \in \mathbb{R}^n.}$$

2. Chaque individu est décrit par k variables, formant un vecteur de dimension k, appelé *vecteur individu*. On la note par

$$\boxed{e_i = (x_{i1}, \ldots, x_{ik})^t \in \mathbb{R}^k.}$$

3. On note par x_{ij} l'observation du caractère x_j sur l'individu e_i.

1.2.2 Poids

Le poids attribué à chaque individu exprime l'importance que l'on désire lui accorder dans l'étude, il arrive parfois que l'on affecte des poids différents à chaque individu. On utilise alors une matrice diagonale :

$$p = \begin{pmatrix} p_1 & 0 & & 0 \\ 0 & p_2 & & 0 \\ & & \ddots & \\ 0 & & & p_n \end{pmatrix}.$$

– $\sum_{i=1}^{n} p_i = 1$, avec $0 \leq p_i \leq 1$.

– On général, le poids est le même pour tous les individus, donc $p_i = \frac{1}{n} I_n$[1].

1.2.3 Matrice de variance-covariance et matrice de corrélation

1.2.3.1 Matrice de variance-covariance

La matrice de variance-covariance ou simplement matrice de covariance d'un vecteur de k variables aléatoires $\vec{X} = (X_1, X_2, ..., X_k)^t$, dont chacune a une variance finie, est une matrice carrée définie comme suit

$$a_{ij} = Cov\left(X_i, X_j\right).$$

Soit X_c la matrice des données centrées définie par

$$X_c = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & & x_{2k} - \bar{x}_k \\ & & & \\ & & \ddots & \\ & & & \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & & x_{nk} - \bar{x}_k \end{pmatrix} \in M_{n\times k}\left(\mathbb{R}\right),$$

où $\bar{x}_j = \sum_{i=1}^{n} p_i x_{ij}$.

[1] On note I_n la matrice indicatrice définie comme suit

$$I_n = \begin{pmatrix} 1 & 0 & & 0 \\ 0 & 1 & & 0 \\ & & \ddots & \\ & & & \\ 0 & & & 1 \end{pmatrix} \in M_{n,n}\left(\mathbb{R}\right).$$

La matrice de variance-covariance notée V est définie comme suit

$$\begin{aligned} V &= X_c^t p X_c \\ &= \begin{pmatrix} Var(X_1) & Cov(X_1,X_2) & & Cov(X_1,X_k) \\ Cov(X_2,X_1) & Var(X_2) & & Cov(X_2,X_k) \\ & & \ddots & \\ Cov(X_k,X_1) & Cov(X_k,X_2) & & Var(X_k) \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1^2 & \sigma_{12} & & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & & \sigma_{2k} \\ & & \ddots & \\ \sigma_{k1} & \sigma_{k2} & & \sigma_k^2 \end{pmatrix}, \end{aligned}$$

où

– $Cov(X_l, X_j) = \sum_{i=1}^{n} p_i (x_{ij} - \bar{x}_j)(x_{il} - \bar{x}_l) = X_{cj}^t p X_{cl}$.

– $Var(X_j) = Cov(X_j, X_j) = \sum_{i=1}^{n} p_i (x_{ij} - \bar{x}_j)^2 = \sigma_j^2$.

Proof. On montre que $Cov(X_l, X_j) = X_{cj}^t p X_{cl}$

$$\begin{aligned} X_{cj}^t p X_{cl} &= \begin{pmatrix} x_{11} - \bar{x}_1 & & x_{n1} - \bar{x}_1 \\ & \ddots & \\ x_{1k} - \bar{x}_k & & x_{nk} - \bar{x}_k \end{pmatrix} \begin{pmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_n \end{pmatrix} \begin{pmatrix} x_{11} - \bar{x}_1 & & x_{1k} - \bar{x}_k \\ & \ddots & \\ x_{n1} - \bar{x}_1 & & x_{nk} - \bar{x}_k \end{pmatrix} \\ &= \begin{pmatrix} p_1 (x_{11} - \bar{x}_1) & & p_n (x_{n1} - \bar{x}_1) \\ & \ddots & \\ p_1 (x_{1k} - \bar{x}_k) & & p_n (x_{nk} - \bar{x}_k) \end{pmatrix} \begin{pmatrix} x_{11} - \bar{x}_1 & & x_{1k} - \bar{x}_k \\ & \ddots & \\ x_{n1} - \bar{x}_1 & & x_{nk} - \bar{x}_k \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^{n} p_i (x_{i1} - \bar{x}_1)^2 & & \sum_{i=1}^{n} p_i (x_{i1} - \bar{x}_1)(x_{ik} - \bar{x}_k) \\ & \ddots & \\ \sum_{i=1}^{n} p_i (x_{ik} - \bar{x}_k)(x_{i1} - \bar{x}_1) & & \sum_{i=1}^{n} p_i (x_{ik} - \bar{x}_k)^2 \end{pmatrix}. \end{aligned}$$

Alors

$$X_{cj}^{t} p X_{cl} = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & & Cov(X_1, X_k) \\ Cov(X_2, X_1) & Var(X_2) & & Cov(X_2, X_k) \\ & & \ddots & \\ Cov(X_k, X_1) & Cov(X_k, X_2) & & Var(X_k) \end{pmatrix}$$
$$= Cov(X_l, X_j).$$

■

Remarque 1.2.2

1. La matrice est symétrique, étant donné la propriété que $Cov(X, Y) = Cov(Y, X)$.
2. Les éléments de sa diagonale représentent la variance de chaque variable.
3. Les éléments en dehors de la diagonale représentent la covariance entre les variables i et j quand $i \neq j$.
4. Ses valeurs propres sont positives ou nulles. Lorsqu'il n'existe aucune relation entre les composantes du vecteur aléatoire, la matrice V est à valeurs propres strictement positives (elle est définie positive).

1.2.3.2 Matrice de corrélation

La matrice de corrélation d'un vecteur de k variables aléatoires $\vec{X}$ est une matrice carrée dont le terme générique est donné par

$$r_{ij} = Cor(X_i, X_j) = \frac{Cov(X_i, X_j)}{\sigma_i \sigma_j}.$$

On a la matrice diagonale de la racine de la variance (écart type)

$$D_\sigma = \begin{pmatrix} \sigma_1 & 0 & & 0 \\ 0 & \sigma_2 & & 0 \\ & & \ddots & \\ 0 & & & \sigma_k \end{pmatrix}.$$

Soit X_r la matrice centrée réduite associée à X définie par

$$\begin{aligned} X_r &= X_c D_\sigma^{-1} \\ &= \begin{pmatrix} \frac{x_{11}-\bar{x}_1}{\sigma_1} & \frac{x_{12}-\bar{x}_2}{\sigma_2} & & \frac{x_{1k}-\bar{x}_k}{\sigma_k} \\ \frac{x_{21}-\bar{x}_1}{\sigma_1} & \frac{x_{22}-\bar{x}_2}{\sigma_2} & & \frac{x_{2k}-\bar{x}_k}{\sigma_k} \\ & & \ddots & \\ \frac{x_{n1}-\bar{x}_1}{\sigma_1} & \frac{x_{n2}-\bar{x}_2}{\sigma_2} & & \frac{x_{nk}-\bar{x}_k}{\sigma_k} \end{pmatrix}. \end{aligned}$$

où D_σ^{-1} est la matrice inverse de D_σ.

La matrice de corrélation notée R est définie comme suit

$$\begin{aligned} R &= X_r^t p X_r \\ &= D_\sigma^{-1} V D_\sigma^{-1} \\ &= \begin{pmatrix} 1 & r_{12} & & r_{1k} \\ r_{21} & 1 & & r_{2k} \\ & & \ddots & \\ r_{k1} & r_{k2} & & 1 \end{pmatrix}. \end{aligned}$$

Les termes diagonaux de cette matrice sont égaux à 1, elle est symétrique, semi-définie positive et ses valeurs propres sont positives ou nulle.

Proof. On montre que $R = D_\sigma^{-1} V D_\sigma^{-1}$

$$\begin{aligned} R &= X_r^t p X_r \\ &= \left(X_c D_\sigma^{-1}\right)^t p \left(X_c D_\sigma^{-1}\right) \\ &= \left(D_\sigma^{-1} X_c^t\right) p \left(X_c D_\sigma^{-1}\right) \\ &= D_\sigma^{-1} \left(X_c^t p X_c\right) D_\sigma^{-1} \\ &= D_\sigma^{-1} V D_\sigma^{-1}. \end{aligned}$$

Donc $R = D_\sigma^{-1} V D_\sigma^{-1}$. ■

1.3 Mesures d'éloignement

Pour regrouper des observations en groupes homogènes, il faut tout d'abord avoir une définition de ce que sont des observations similaires ou des observations dissimilaires. Il faut donc être en mesure de quantifier la similarité ou la distance ou la dissimilarité entre deux observations. Cette première étape peut parfois être la plus difficile de tout le processus de classification, mais elle est essentielle.

1.3.1 Distance et dissimilarité

On note $\Omega = \{e_1, \ldots, e_n\}$ l'ensemble des individus.

1.3.1.1 Dissimilarité

Une dissimilarité est une application d de $\Omega \times \Omega$ dans $\mathbb{R}_+$ vérifiant

– $\forall (e_1, e_2) \in \Omega^2 : d(e_1, e_2) \geq 0$.

– $\forall (e_1, e_2) \in \Omega^2 : d(e_1, e_2) = 0$ si et seulement si $e_1 = e_2$.

– $\forall (e_1, e_2) \in \Omega^2 : d(e_1, e_2) = d(e_2, e_1)$.

1.3.1.2 Distance

On peut donc aborder le problème de la ressemblance entre individus par le biais de la notion de distance. On appelle distance notée d sur un ensemble Ω toute application $d : \Omega \times \Omega$ dans $\mathbb{R}_+$ telle que

– $\forall (e_1, e_2) \in \Omega^2 : d(e_1, e_2) = 0$ si et seulement si $e_1 = e_2$.

– $\forall (e_1, e_2) \in \Omega^2 : d(e_1, e_2) = d(e_2, e_1)$.

– $\forall (e_1, e_2, e_3) \in \Omega^3 : d(e_1, e_2) \leq d(e_1, e_3) + d(e_3, e_2)$.

Remarque 1.3.1

1. Si Ω est un fini, la distance peut être normée.
2. Une distance est une dissimilarité mais l'inverse n'est pas vrai.
3. On utilise souvent une distance pour évaluer une dissimilarité.

1.3.1.3 Mesures de distance

La classification hiérarchique utilise des mesures de dissemblance ou de distance entre les objets pour former des classes, il existe plusieurs types de distances parmi lesquels on trouve :

a/ Distance Euclidienne :

C'est une mesure possible de la ressemblance, c'est probablement le type de distance le plus couramment utilisé**.** Elle est donnée par la formule ci-dessous

$$\boxed{\forall i, i' = 1, \ldots, n : d(e_i, e_{i'}) = \sqrt{\sum_{j=1}^{k} (x_{ij} - x_{i'j})^2}.}$$

b/ Distance de Manhattan (métrique city block) :

Cette distance est simplement la somme des différences entre les dimensions**.** Elle est donnée par la formule ci-dessous

$$\boxed{\forall i, i' = 1, \ldots, n : d(e_i, e_{i'}) = \sum_{j=1}^{k} |x_{ij} - x_{i'j}|.}$$

c/ Distance Tchebyshev :

C'est la distance entre deux points donnée par la différence maximale entre leurs coordonnées sur une dimension**.** Elle est donnée par la formule ci-dessous

$$\boxed{\forall i, i' = 1, \ldots, n : d(e_i, e_i) = \max_{j=1,\ldots,k} |x_{ij} - x_{i'j}|.}$$

d/ Distance de Minkowski :

C'est une métrique dans un espace vectoriel normé qui peut être considérée comme une généralisation à la fois de la distance euclidienne et de la distance de Manhattan, à été proposée par *[12]***.** Elle est donnée par la formule ci-dessous

$$\boxed{\forall i, i' = 1, \ldots, n : d(e_i, e_{i'}) = \left(\sum_{j=1}^{k} |x_{ij} - x_{i'j}|^q \right)^{\frac{1}{q}} ; q \geq 1.}$$

Remarque 1.3.2

1. La distance de Minkowski est généralement utilisée lorsque q est égal à 1 ou à 2, ce qui correspondent respectivement à la distance de Manhattan et à la distance euclidienne.
2. Dans le cas limite d'atteindre l'infini, on obtient la distance de Tchebyshev :

$$\forall i, i' = 1, \ldots, n : d(e_i, e_{i'}) = \lim_{q \to +\infty} \left(\sum_{j=1}^{k} |x_{ij} - x_{i'j}|^q \right)^{\frac{1}{q}} = \max_{j=1,\ldots,k} |x_{ij} - x_{ij}|.$$

e/ Distance Mahalanobis [11] :

Elle est donnée par la formule ci-dessous

$$\boxed{\forall i, i' = 1, \ldots, n : d(e_i, e_{i'}) = \sqrt{(x_i - x_{i'})^t V^{-1} (x_i - x_{i'})},}$$

où V^{-1} est l'inverse de la matrice de covariance.

Remarque 1.3.3

1. Si la matrice de covariance est la matrice identité, cette distance est simplement la distance euclidienne.
2. Si la matrice de covariance est diagonale, on obtient la distance euclidienne normalisée suivante
$$\forall i, i' = 1, \ldots, n : d(e_i, e_{i'}) = \sqrt{\sum_{j=1}^{k} \frac{(x_{ij} - x_{i'j})^2}{\sigma_j^2}}.$$

f/ Distance de Canberra :

La distance de Canberra est une mesure numérique de la distance entre paires de points dans un espace vectoriel, introduite par [6]et *[7]*. Elle est donnée par la formule ci-dessous

$$\boxed{\forall i, i' = 1, \ldots, n : d(e_i, e_i) = \sum_{j=1}^{k} \frac{x_{ij} - x_{i'j}}{x_{ij} + x_{i'j}}.}$$

1.3.2 Similarité entre objets décrits par des variables binaires

Dans le cas où n individus sont décrits par la présence ou l'absence de η caractéristiques. De nombreux indices de similarité ont été proposés qui combinent de diverses manières, les quatre nombres suivants associés à un couple d'individus :

$j \diagdown i$	1	0	
1	α	γ	$\alpha+\gamma$
0	β	δ	$\beta+\delta$
	$\alpha+\beta$	$\gamma+\delta$	η

TAB. 1.1 – Tableau des caractéres pour les données dichtomiques.

- α représente le nombre de caractéristiques communes à i et j.
- β représente le nombre de caractéristiques possédées par i mais pas par j.
- γ représente le nombre de caractéristiques possédées par j mais pas par i.
- δ représente le nombre de caractéristiques que ne possédées ni i ni j.
- bien sûr $\alpha+\beta+\gamma+\delta=\eta$.

Les indices de ressemblance les plus courants sont :

- **Coefficient d'appariement (matching) simple :**

$$\boxed{d(i,j)=\frac{\beta+\gamma}{\eta}.}$$

- **Coefficient de Russel et Rao [13] :**

$$\boxed{d(i,j)=\frac{\alpha}{\eta}.}$$

- **Coefficient de Jaccard [9] :**

$$\boxed{d(i,j)=\frac{\alpha}{\alpha+\beta+\gamma}.}$$

– **Coefficient de Dice [3] :**

$$\boxed{d(i,j) = \frac{2\alpha}{2\alpha + \beta + \gamma}.}$$

– **Coefficient de Sokal et Sneath [14] :**

$$\boxed{d(i,j) = \frac{\alpha}{\alpha + 2(\beta + \gamma)}.}$$

Exemple 1.3.1 *Soit deux objets* O_i *et* O_j *tels que*
$O_i = (1,0,1,0,0,1,1)$ *et* $O_j = (0,0,1,1,1,1,0)$. *On a*

- Coefficient de Dice : $d(i,j) = 4/(4+2+1) = 4/7$.

- Coefficient de Jaccard : $d(i,j) = 2/5$.

1.4 Méthodes hiérarchiques

La classification hiérarchique consiste à effectuer une suite de regroupements en classes de moins en moins en fines en agrégeant à chaque étape les objets ou les groupes d'objets les plus proches. Elle fournit ainsi un ensemble de partitions de l'ensemble d'objets [**1**]. Cette approche utilise la notion de distance, qui permet de refléter l'homogénéité ou l'hétérogénéité des classes. Ainsi, on considère qu'un élément appartient à une classe s'il est plus proche de cette classe que de toutes les autres.

1.4.1 Hiérarchie

Une hiérarchie est un ensemble de partitions emboitées. On la représente en général par un arbre hiérarchique appelé dendrogramme où les objets sont à la base l'ensemble tout entier au sommet.

Définition 1.4.1 (***Hiérarchie***)*On appelle hiérarchie* **H** *de* Ω *tout ensemble de parties de* Ω *vérifiant les propriétés suivantes :*

- $\emptyset \notin \mathbf{H}$.
- $\Omega \in \mathbf{H}$.
- Pour tout élément de Ω, la hiérarchie contient tous les singletons i.e.

$$\forall e \in \Omega : \{e\} \in \mathbf{H}.$$

- Soient C_1 et C_2 deux parties quelconque de **H**, on a $C_1 \cap C_2 \in \{\emptyset, C_1, C_2\}$ autrement dit C_1 et C_2 sont soit disjointes, alors l'une est incluse dans l'autre.

Définition 1.4.2 ***(Hiérarchie de partie indicée)*** *C'est une hiérarchie de parties H à laquelle est associée une échelle d'indices qui satisfait les propriétés suivantes.*

Il existe une application i positive définie sur **H** telle que

- $\forall i(A) \geq 0 : A \subseteq B \Longrightarrow i(A) \leq i(B)$.
- $\forall e \in \Omega : i(e) = 0$.

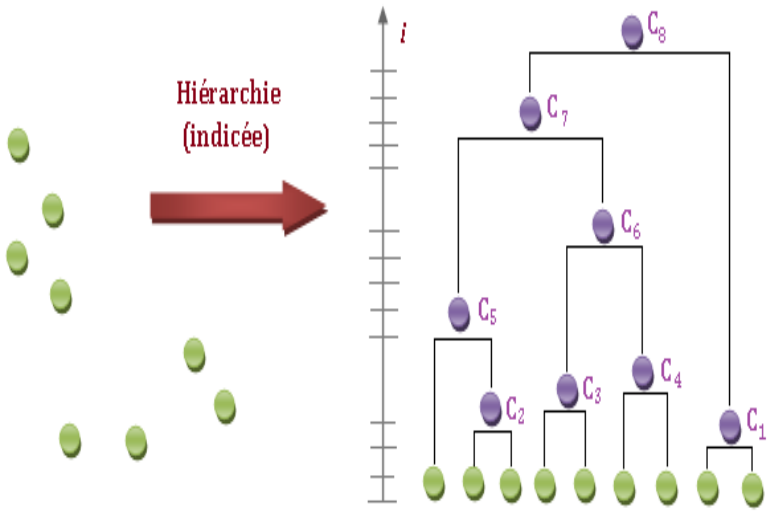

FIG. 1.1 – Hiérarchie indicée.

1.4.2 Classification ascendante hiérarchique (CAH)

La classification ascendante hiérarchique est la méthode la plus utilisée. Elle consiste à construire une succession de partitions emboîtées par regroupement successifs des observations (objets) en classes de moins en moins fines, jusqu'à l'obtention d'une seule classe contenant tous les objets.

1.4.2.1 Principe de la CAH

Initialement, chaque individu forme une classe, soit n classes. À chaque étape, on fusionne deux classes, réduisant ainsi le nombre de classes. Les deux classes choisies pour être fusionnées sont celles qui sont les plus "proches", en d'autres termes, celles dont la dissimilarité entre elles sont minimale, cette valeur de dissimilarité est appelée indice d'agrégation. Comme on rassemble d'abord les individus les plus proches, la première itération a un indice d'agrégation faible, mais celui-ci va croître d'itération en itération.

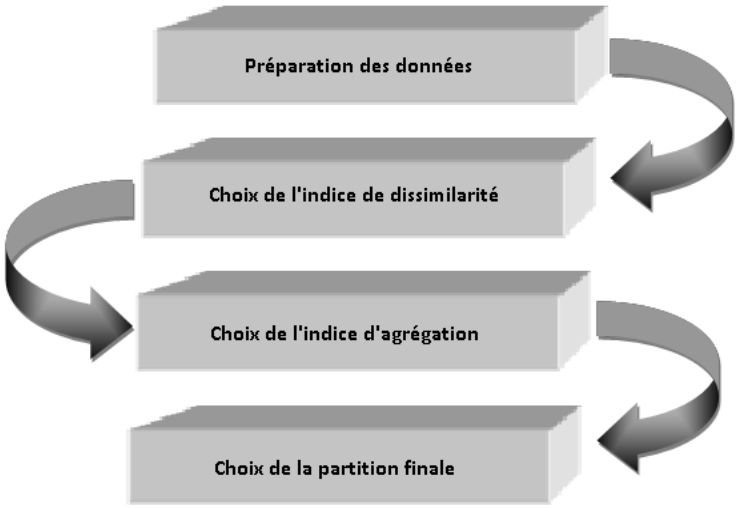

FIG. 1.2 – Etapes de la CAH.

1.4.2.2 Matrice de distance

La CAH consiste à calculer une matrice exprimant les distances mutuelles entre les points à classer, appelée matrice des distances ou des dissimilarités, c'est une matrice de taille $n \times n$ représentant l'espacement d'un ensemble de n points. On note par D la matrice de distance.

$$D = \begin{pmatrix} 0 & d(e_1, e_2) & & d(e_1, e_n) \\ d(e_2, e_1) & 0 & & d(e_2, e_n) \\ & & \ddots & \\ d(e_n, e_1) & d(e_n, e_2) & & 0 \end{pmatrix} \in M_{(n\times n)}(\mathbb{R}).$$

Remarque 1.4.1

1. Tous les éléments sur la diagonale de D sont nuls (on parle alors de matrice creuse).

2. D est symétrique.

3. Le nombre de valeurs uniques (distinctes) non nulles d'une matrice de distance euclidienne de taille $n \times n$ est borné supérieurement par $n(n-1)/2$ en raison de la symétrie et du fait d'être creuse.

4. Généralement pour cette matrice on l'utilise la distance euclidienne.

1.4.2.3 Critères d'agrégation

Dans la CAH, il existe plusieurs méthodes d'agrégation possibles entre deux classes. Le choix de la méthode dépend de la problématique que l'on se pose. En ce qui concerne la distance entre deux individus, la distance euclidienne est la plus utilisée en général.

Parmi les méthodes d'agrégation existantes pour chaque couple de classes, on peut citer :

– **Méthode de la liaison simple (ou critère du saut minimal)**

Soient $D(C_1, C_2)$ la distance entre les groupes C_1 et C_2. L'agrégation entre deux groupes C_1 et C_2 correspond au minimum des distances entre un élément du groupe C_1 et un élément du groupe C_2.

$$D(C_1, C_2) = \min(\{d(x, y)\}; x \in C_1, y \in C_2).$$

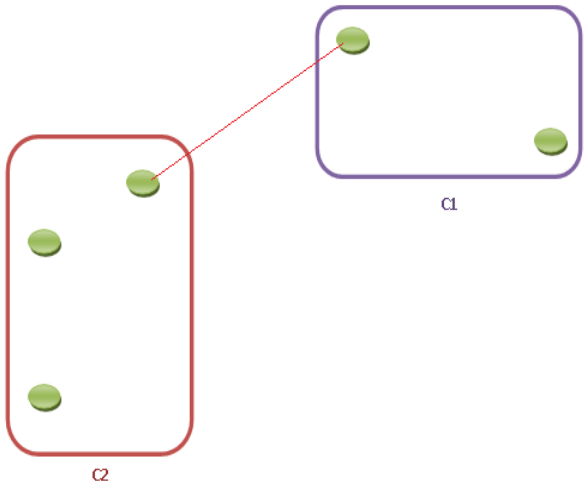

FIG. 1.3 – Saut minimal.

– **Méthode de la liaison complète (ou critère du saut maximal)**

L'agrégation entre deux groupes C_1 et C_2 correspond au maximum des distances entre un élément du groupe C_1 et un élément du groupe C_2.

$$D(C_1, C_2) = \max(\{d(x, y)\}; x \in C_1, y \in C_2).$$

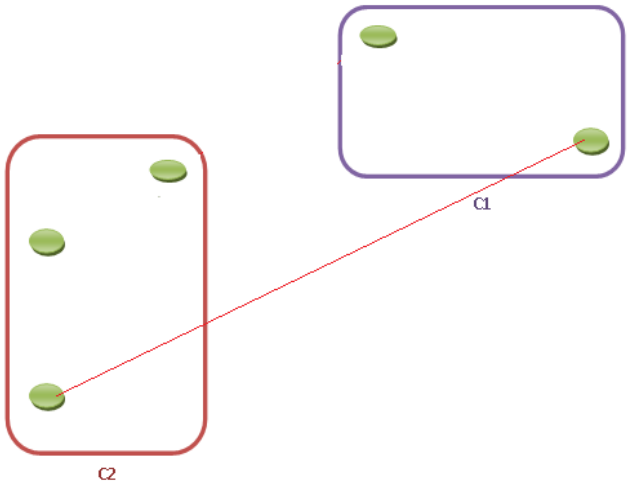

FIG. 1.4 – Saut maximal.

– **Méthode de la liaison moyenne**

Elle se détermine en calculant toutes les distances possibles entre les observations respectives aux deux groupes étudiés et en déterminant la moyenne de ces distances. Les deux groupes dont la distance moyenne est la plus petite sont ceux qui seront fusionnés pour l'itération en cours.

$$D\left(C_1, C_2\right) = \frac{1}{n_{C_1} n_{C_2}} \sum_{x \in C_1} \sum_{y \in C_2} d\left(x, y\right).$$

où n_{C_1} et n_{C_2} est le nombre d'objets dans chacun des deux groupes C_1 et C_2 respectivement.

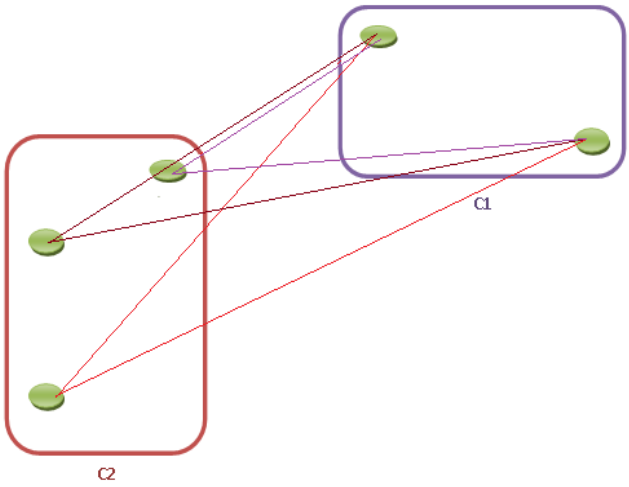

FIG. 1.5 – Liaison moyenne.

– **Méthode de Ward**

Ce critère ne s'applique que si on est muni d'un espace euclidien. La dissimilarité entre deux individus doit être égale à la moitié du carré de la distance euclidienne d. Le critère de Ward consiste à choisir à chaque étape le regroupement de classes tel que l'augmentation de l'inertie intra-classe soit minimale. Ce critère proposé notamment par *[15]* est considéré comme suit

$$D = \frac{n_{C_1} n_{C_2}}{n_{C_1} + n_{C_2}} d^2 \left(g_{C_1}, g_{C_2}\right).$$

– **Méthode de centre de gravité**

La distance entre deux groupes G_1et G_2 est définie par la distance entre leurs centres de gravité

$$D = d \left(g_{C_1}, g_{C_2}\right).$$

avec

g_{C_1} et g_{C_2} le centre de gravité de C_1 et C_2 respectivement.

Inertie

On a

Poids de C_j : $q_k = \sum_{e_i \in C_j} p_i$ pour $j = 1, \ldots, k$.

Centre de gravité de C_j : $g_j = \frac{1}{q_k} \sum_{e_i \in C_j} p_i e_i$ pour $i = 1, \ldots, n$.

– Inertie totale : L'inertie totale du nuage des n individus notée I_{tot} est donnée par

$$I_{tot} = \sum_{i=1}^{n} p_i d_M^2 (e_i, g).$$

où g le centre de gravité du nuage des individus.

– Inertie inter-classe : L'inertie inter-classe de la partition P_k notée I_{inter} est l'inertie des centres de gravité des classes pondérées par q_k. Elle est donnée par

$$I_{inter} = \sum_{j=1}^{k} q_k d_M^2 (g_j, g).$$

– Inertie intra-classe : L'inertie intra-classe de la partition P_k notée I_{intra} est la somme des inerties des classes. Elle est donnée par

$$I_{intra} = \sum_{j=1}^{k} \sum_{e_i \in C_j} P_i d_M^2 (e_i, g_j).$$

– **Décomposition de Huygens :** Pour tout P_k partition de Ω, on a

$$I_{tot} = I_{inter} + I_{intra}.$$

On constate que minimiser l'inertie intra-classe est équivalent à maximiser l'inertie inter-classe. Cette décomposition est illustrée par la figure 1.6.

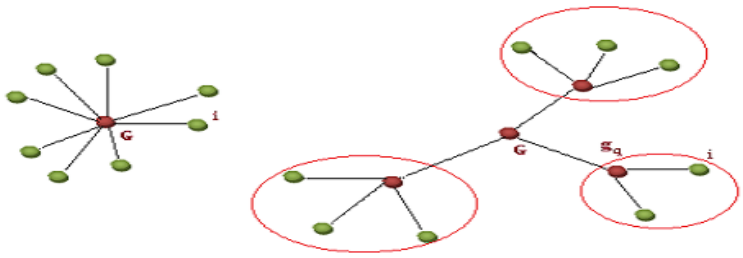

FIG. 1.6 – Décomposition de Huygens.

Remarque 1.4.2 *La méthode de Ward est la méthode la plus souvent utilisée car elle a été conçue dans un objectif de maximisation de l'inertie inter-classe et de minimisation de l'inertie intra-classe.*

1.4.2.4 Avantages de la CAH

- Aide au choix du nombre de groupes : La plupart des méthodes de clustering demandent à l'utilisateur de choisir le nombre de groupes qu'il souhaite créer. Ce n'est pas le cas de la CAH qui va calculer toutes les combinaisons possibles. Elle les représente ensuite via un dendrogramme qui permettra au Data Scientiste de choisir le nombre de clusters le plus adapté à ses données et à son objectif.
- Facile à utiliser.
- Détecte des classes de forme diverse ou des centres de classes.
- Permet d'utiliser différents types de variables.

1.4.2.5 Inconvénients de la CAH

- Pas adapté aux grands volumes de données.
- Exigeante en termes de temps de calcul.
- Une fois que deux individus sont groupés, ils ne seront jamais séparés.

1.4.3 Classification descendante hiérarchique (CDH)

On a vu dans la précédente méthode que la classification hiérarchique ascendante, essaie d'optimiser un seul critère à la fois, ceci engendre uniquement une séparation (méthode du lien simple) ou une homogénéité (méthode du lien complet) optimale des classes, ce qui risque de donner naissance à l'effet de chaînage (deux entités très dissimilaires appartenant aux points extrêmes d'une longue chaîne, peuvent appartenir à la même classe) ou l'effet de dissection (deux entités très similaires peuvent être dans deux classes différentes).
Pour palier à ces problèmes, il existe des algorithmes divis ifs de la classification hiérarchique descendante. Ces méthodes ont eu moins de succès que les premières. Les algorithmes divis ifs commencent par former une seule classe qui englobe tous les objets. Par la suite, ils choisissent une classe de la partition en cours selon un premier critère local. Ils procèdent ensuite à une bipartition successive selon un deuxième critère local des classes choisies. Cette bipartition continue jusqu'à ce que toutes les entités soient affectées à différentes classes.
Parmi les algorithmes les plus anciens, l'algorithme de **Williams** et **Lambert** [**16**] qui divise la plus grande classe en deux classes, l'algorithme de **Hubert** qui propose de diviser la classe de plus grand diamètre et l'algorithme **TSVQ** (Tree Structured Vector Quantization) qui a été proposé par *[5]*.

1.4.3.1 Inconvénients de la CDH

Les résultats sont en général grossiers, les niveaux des nœuds de la hiérarchie ne sont plus définis que par l'ordre dans lequel ils apparaissent.

1.4.3.2 Avantages de la CDH

Malgré ses nombreux inconvénients la classification descendante présente quelques avantages :

- Une des propriétés intéressantes de cet algorithme hiérarchique descendant est la stabilité de leurs résultats.
- La méthode de la classification hiérarchique descendante ne nécessite pas l'utilisation d'un seuil arbitraire pour la formation des classes qui peut éventuellement mener la recherche d'une classification d'un ensemble de données à une fausse direction.

1.4.4 Classification des variables

L'objectif de la classification des variables est d'obtenir des classes de variables liées. Il existe deux principaux types de variables, chacune ayant deux sous-types selon son système de classification.

1.4.4.1 Variables quantitatives

Une variable quantitative dis aussi numérique est une variable qui reflète une notion de grandeur, c'est-à-dire si les valeurs qu'elle peut prendre sont des nombres. Une variable quantitative représente donc une mesure numérique.
Les variables quantitatives sont divisées en deux types : discrètes et continues. La différence est expliquée ci-dessous :

– **Variables discrètes :** Dont les modalités sont des nombres appartenant à un ensemble fini de valeurs, par exemple le nombre de victimes lors du sinistre.
– **Variables continues :** Dont les modalités sont des nombres appartenant un intervalle réel par exemple le kilométrage d'un véhicule ou le montant d'un sinistre.

1.4.4.2 Variables qualitatives

Une variable qualitative dis aussi catégorielles est une variable qui n'est pas numérique et dont les valeurs s'inscrivent dans des catégories. Les variables qualitatives peuvent être divisées en deux types :

- **Variables ordinales :** Les valeurs reflètent un ordre, rien plus. Des exemples de variables catégorielles ordinales comprennent les notes académiques i.e. A, B et C, la taille des vêtements i.e. petite, moyenne, grande et très grande, etc.
- **Variables nominales :** Les valeurs sont juste des noms différents. Par exemple les codes postaux, les couleurs, le sexe, etc. La valeur est prise dans une liste finie.

Chapitre 2

Classification non hiérarchique et application

Contrairement à la méthode hiérarchique Une méthode non hiérarchique génère une classification en partitionnant un ensemble de données, donnant un ensemble de groupes généralement non chevauchants n'ayant pas de relations hiérarchiques entre eux. Une évaluation systématique de toutes les partitions possibles est tout à fait irréalisable, de nombreuses heuristiques différentes ont ainsi été décrites pour permettre l'identification de bonnes partitions. Les méthodes non hiérarchiques sont généralement beaucoup moins exigeantes en ressources de calcul que les méthodes hiérarchiques.

L'algorithme général d'une méthode de partitionnement comprend les étapes suivantes :

Etape1 : Déterminer le nombre de groupes.

Etape2 : Initialiser les centres des groupes.

Etape3 : Partitionner l'ensemble des données.

Etape4 : Calculer les centres des groupes.

Etape5 : Si le partitionnement est inchangé (ou l'algorithme a convergé) arrêt, sinon revenir à l'étape 3.

La classification non hiérarchique d'un ensemble d'individus peut être divisée en trois grandes sous-familles : K-means [10], centres mobiles [4] et Nuées dynamiques [2].

2.1 Classification non hiérarchique

2.1.1 Principe de la classification non hiérarchique

Le principe général de cette méthode est permet de subdiviser l'ensemble des individus en un certain nombre de classes donnant un ensemble de groupes n'ayant pas de relations hiérarchiques entre eux, telle que chaque classe doit contenir au moins un individu et chaque individu doit appartenir à une classe unique. Le critère généralement utilisé pour savoir si le résultat de diviser est bon ou non est que les objets d'une même classe doivent être très proches les uns des autres et très éloignés des autres classes.
Ces méthodes de partitionnement sont basées sur une distance ou un indice de similarité entre objets à classer.

Définition 2.1.1 ***(partition)*** *Une partition P_k de Ω en k classes est un ensemble*

$\{C_1, \ldots, C_k\}$ de classes non vides, vérifiant

- $\forall i, j = 1, \ldots, k; i \neq j : C_i \cap C_j = \emptyset$.
- $\bigcup_{i=1}^{k} C_i = \Omega$.

Exemple 2.1.1 *Soit* $\Omega = \{1, \ldots, 8\}$ *et* $P_4 = \{C_1, C_2, C_3, C_4\}$ *avec* $C_1 = \{1, 2\}$,

$C_2 = \{5, 7\}, C_3 = \{3\}, C_4 = \{4, 6, 8\}$ est une partition de quatre classes de Ω.

2.1.2 Différents Algorithmes

Les algorithmes de classification non hiérarchique sont divisés en trois grandes sous-familles.

2.1.2.1 Méthode de K-means (Mac Queen)

L'algorithme de K-means ou K-means clustering en anglais a été introduit par **[10]**, c'est la méthode de partitionnement la plus connue et la plus utilisée dans divers domaines d'application scientifiques et industrielles. Dans ce type, la classe est représentée par son centre de gravité, cette méthode a pour but de diviser des observations en k partitions dans les quelles chaque observation appartient à la partition avec la moyenne la plus proche.
Le principe de cette méthode est très simple, il est basé sur la centre de gravité, il attribue chaque individu dans la classe s'il est très proche de son centre de gravité. Le centre est la moyenne de tous les individus dans la classe.

2.1.2.1.1 Algorithme de K-means

L'algorithme de K-means se déroule de la façon suivante :

Etape 1 : Choisir k individu au hasard comme centre des classes initiales.

Etape 2 : Affecter chaque individu vers centre de classe la plus proche.

Etape 3 : Recalculer les centres de gravité des nouvelles classes.

Etape 4 : Répéter l'étape 2 et 3 jusqu'à convergence, i.e. lorsque les positions du centre de gravité ne changent plus.

Etape 5 : Finalement éditer la partition obtenue.

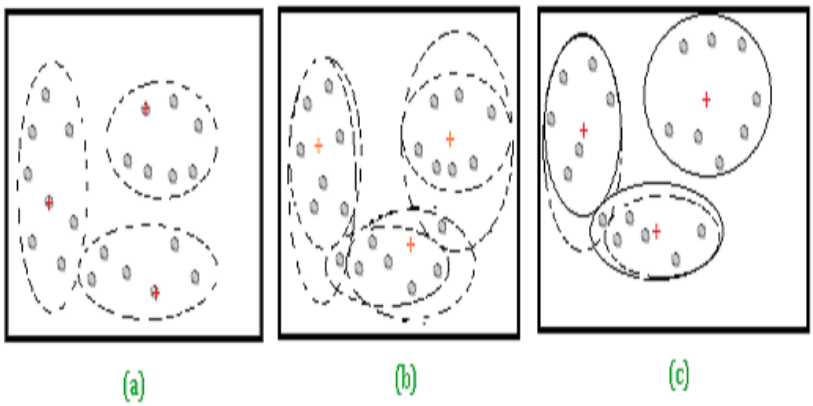

FIG. 2.1 – Partitionnement basé sur la méthode de K-means.

2.1.2.1.2 Avantages de K-means

- Simple, rapide et facile à comprendre.
- Flexible de manière à s'adapter aux divers changements des données.
- Faible coût de calcul : comparée à l'utilisation d'autres méthodes de classification une technique de classification de K-means est rapide et efficace en termes de coût de calcul.
- Applicable à des données de grandes tailles et à tout type de données.

2.1.2.1.3 Inconvénients de K-means

- Il ne peut être exécuté qu'avec des données numériques.
- La répartition en classes dépend du choix initial des centres (faire tourner l'algorithme plusieurs fois pour identifier des formes fortes).

2.1.2.2 Méthode du centre mobile (Forgy)

La méthode du centre mobile peut être attribuée principalement à **[4],** est une des techniques de partitionnement les plus utilisées. Elle s'applique lorsqu'on sait à l'avance combien de classes on veut obtenir. L'objectif de la méthode est de partitionner en différentes classes des individus pour lesquels on dispose certaines mesures. Le but de cette méthode

est de regrouper autant que possible les individus les plus semblables tout en séparant les classes le mieux possible les unes des autres. Cette méthode est comme dans le cas de la classification hiérarchique ascendante, i.e. on choisit de procéder de façon automatique en utilisant des mesures de ressemblances et de différences (celle qui est à priori peu visibles).

2.1.2.2.1 Algorithme du centre mobile

Etape 1 : On tire au hasard k individus qu'on considère comme des centres pour les classes initiales.

Etape 2 : On effectue chaque individu au centre le plus proche pour obtenir une partition de k classes $P_0 = \{C_1, C_2, \ldots, C_k\}$.

Etape 3 : On désigne les nouveaux centres de chaque classe de P_0.

Etape 4 : On répète les étapes 2 et 3 jusqu'à ce que deux itérations successives donnent la même partition.

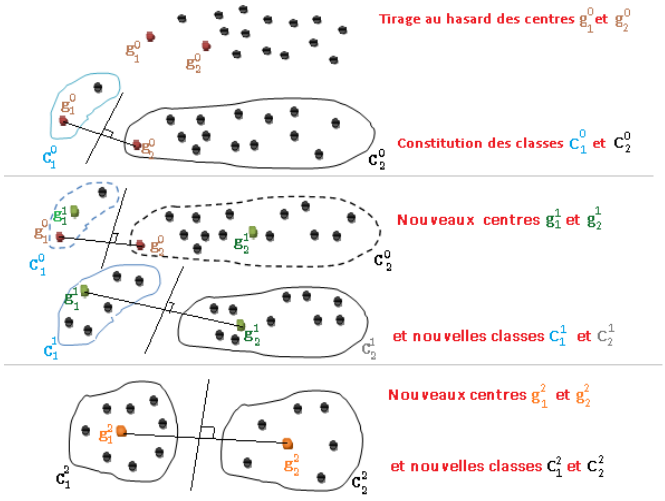

FIG. 2.2 – Méthode du centre mobile.

2.1.2.2.2 Avantages du centre mobile

La méthode du centre mobile comporte plusieurs avantages :

- Simplicité et facilité d'utilisation.
- Compréhensible.
- Suffit souvent de peu d'intérations pour avoir déjà une partition de qualité.
- Son adaptation aux larges bases des données.

2.1.2.2.3 Inconvénients du centre mobile

La méthode du centre mobile comporte des inconvénients, comme par exemple :

- Les nombre de classe doit être fixé au départ.
- Le nombre de classe est un paramètre de l'algorithme, un bon choix du nombre k est nécessaire car un mauvais choix de k produit de mauvais résultat.

2.1.2.3 Méthode du type Nuées dynamiques (Diday)

Une méthode de classification plus générale appelée nuées dynamiques à été proposée par *[2]*. Cette méthode permet une représentation plus générale des classes à travers un noyau formé de points. Elle peut être considérée comme une généralisation de la méthode des centres mobiles. Le principe de cette méthode est très simple, elle consiste à choisir au hasard k noyaux parmi une famille de noyaux telle que chaque noyau contient un sous-ensemble d'individus. Puis chaque point de l'ensemble d'apprentissage est affecté au noyau dont il est le plus proche. On obtient ainsi une partition en k classes dont on calcule les noyaux. On répète le processus précédent avec de nouveaux noyaux jusqu'à ce que la qualité de la partition ne s'améliore plus.

2.1.2.3.1 Algorithme des Nuées dynamiques

Les étapes à suivre sont :

Etape 1 : Choisir arbitrairement k groupes de n points chacun (noyon initiaux) autour duquel on agrège le reste des point (autres points) de E selon le critère du plus proche.

Etape 2 : Déterminer de nouveaux noyaux qui sont les ensembles de n point d'une classe les plus proches des autres points de la classe.

Etape 3 : On refait la même chose avec les nouveaux noyaux jusqu'à convergence de l'algorithme vers la partition finale de E, les noyaux finaux sont appelés les individus typiques d'une classe.

Dans la figure 2.3 explique l'algorithme de cette méthode avec un nombre de classe $k = 2$.

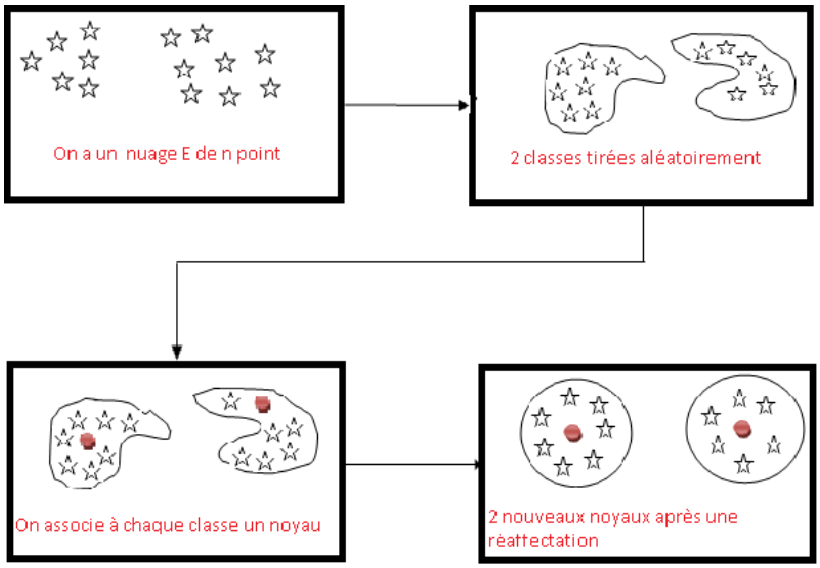

FIG. 2.3 – Algorithme des Nuées dynamiques.

2.1.2.3.2 Avantages des Nuées dynamiques

- Traiter rapidement de grands ensembles d'individus.
- Amélioration continue de la qualité des classes.

- Applicable à des données de grandes tailles.

2.1.2.3.3 Inconvénients des Nuées dynamiques

- Nombre de classe fixé d'avance.
- Fournit une solution dépendante de la configuration initiale et nécessite le choix du nombre de classes.
- Pour comparer l'individu avec les noyaux, cette méthode utilise des distances ce qui est un inconvénient d'établir des métriques.

2.2 Application

On va illustrer la partie théorique par un exemple (programmé sous logiciel R) sur la pandémie du coronavirus appelée aussi Covid-19. Les statistiques prises sont datées à partir des février 2020 jus qu'an fin mai 2021. Ces données sont extraites du site web [8]. Le but de cette partie est de faire une classification automatique de 23 pays arabes (objets) décrits par le nombre de personnes (variables) contaminées, guéries et décédés, afin d'identifier la similarité ou la dissimilarité entre les pays celons leurs nombre de mortalité on de guérison.

Diverses packages sont utilisés dans cette application citons comme exemple : stats, read.csv2, hclust, kmeans.

Charger des données :

```
library(stats)                                          # Importer un tableau.
base=read.csv2("C :/Users/OASIS/Desktop/donn.cvs")      # Tableau des données.
print(base)
fix(base)                                               # Tableau similaire à L'Excel.
```

corona=cbind(base[,2 :4]) # Suppression de la $1^{ière}$ colonne.

dim(corona) # Dimension du tableau.

23 3

names(corona) # Entête du tableau.

"Blessures" "Mortalité" "Guérison"

Statistique déscriptive :

sammary(corona) # Calcul du minimum, moyenne, maximum, mediane, quartile 1 et 3.

	Blessures	Mortalité	Guérison
Min.	1466	115.0	2927
1st Qu.	21981	509.5	14920
Median	217458	1771.0	191475
Mean	275733	4287.4	248252
3rd	397562	7542.5	367641
Max	1201352	16375.0	1116456

pairs(corona) # Graphique-croisement deux à deux.

corona.cr=scale(corona,center=T,scale=T) # Centrage et réduction des données.

d.corona=dist(corona.cr) # Matrice des distances entre les individus.

par(mfrow=c(2,2)) # Partitionner la page en quatre.

Graphe :

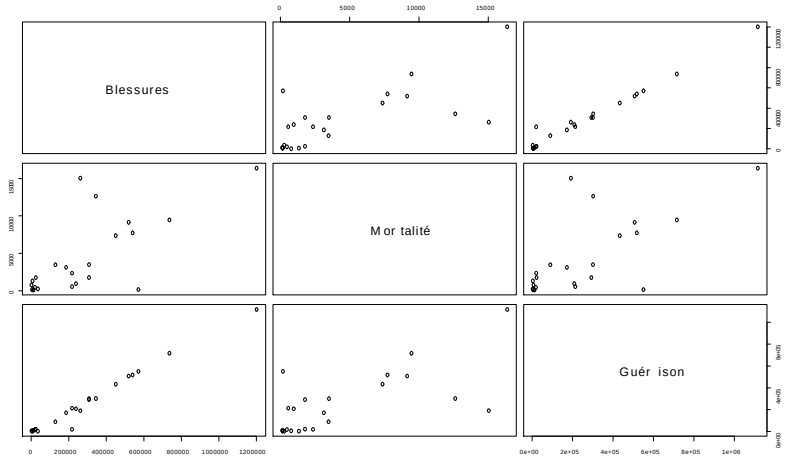

FIG. 2.4 – Croissement deux à deux.

Commentaire :

D'après la figure ci-dessus, on constate que le nombre de "Blessures" est fortement corrélé avec le nombre de personnes guéries, ce qui est évident. Inversement, on remarque une faible corrélation entre le nombre de personnes blessées et le nombre de mortalité.

Classification hiérarchique :

On va présenter les quatre méthodes de la classification hiérarchique vu dans la partie théorique.

Méthode de Ward :

```
cah.ward=hclust(d.corona,method="ward.D2")
plot(cah.ward, hang=-1)              # Dendrogramme de la méthode de Ward.
rect.hclust(cah.ward,k=3)            # Découpage du dendrogramme en trois sous groupes.
dcah.ward=cutree(cah.ward,k=3)       # Découpage numérique des pays en trois sous groupes.
```

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	2	2	2	2	2	2	3	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3

Graphe :

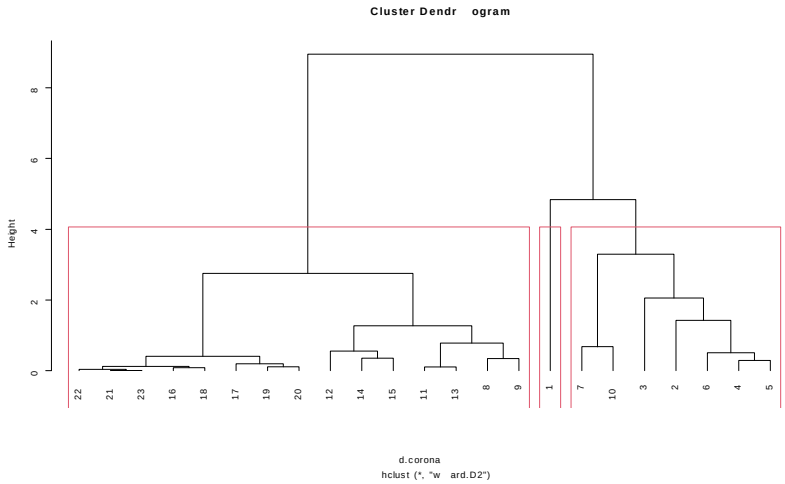

FIG. 2.5 – CAH par la méthode de Ward.

Commentaire :

Le dendogramme suggére un découpage en trois groupes.

$1^{ére}$ **classe :** Elle concerne les pays les moins contaminés et par conséquent, leurs nombre de guérison est assez élevé.

$2^{ème}$ **classe :** L'Irak, le pays le plus contaminé et dont le nombre de mortalité est très élevé.

$3^{ème}$ **classe :** Elle concerne les pays où leurs nombre de contamination est assez faible par rapport à la deuxième classe, mais elle reste quand même supérieure en la comparant avec le première classe.

Méthode de la liaison moyenne :

```
cah.ave=hclust(d.corona,method="ave")
plot(cah.ave, hang =-1)                  # Dendrogramme de la méthode de la liaison moyenne.
rect.hclust(cah.ave,k=3)                 # Découpage en trois sous groupes.
dcah.ave=cutree(cah.ave,k=3)             # Découpage numérique en trois sous groupes.
```

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	2	2	2	2	2	2	3	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3

Graphe :

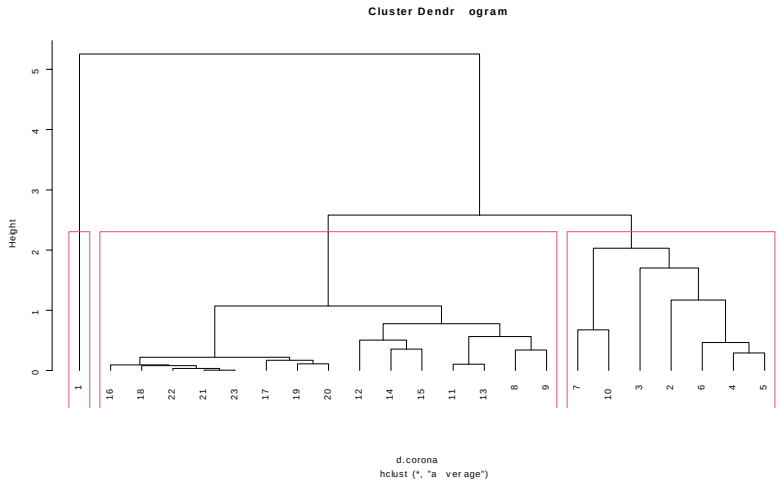

FIG. 2.6 – CAH par la méthode de la liaison moyenne.

Commentaire :

$1^{ére}$ **classe :** L'Irak, le pays le plus contaminé et dont le nombre de mortalité est très élevé.

$2^{ème}$ **classe :** Elle concerne les pays les moins contaminés et par conséquent, leurs nombre de guérison est assez élevé.

$3^{ème}$ **classe :** Elle concerne les pays où leurs nombre de contamination est assez faible par rapport à la première classe, mais elle reste quand même supérieure en la comparant avec la deuxième classe.

Les résultats obtenus par cette méthode sont similaire à ceux de la méthode de ward, mais avec un niveau d'agrégation different.

Méthode du saut maximal :

```
cah.complete=hclust(d.corona, method="complete",members=NULL)
plot(cah.complete, hang=-1)                  # Dendrogramme de la méthode du saut maximal.
rect.hclust(cah.complete,k=3)                # Découpage en trois sous groupes.
dcah.complete=cutree(cah.complete,k=3)       # Découpage numérique en trois sous groupes.
```

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3

Graphe :

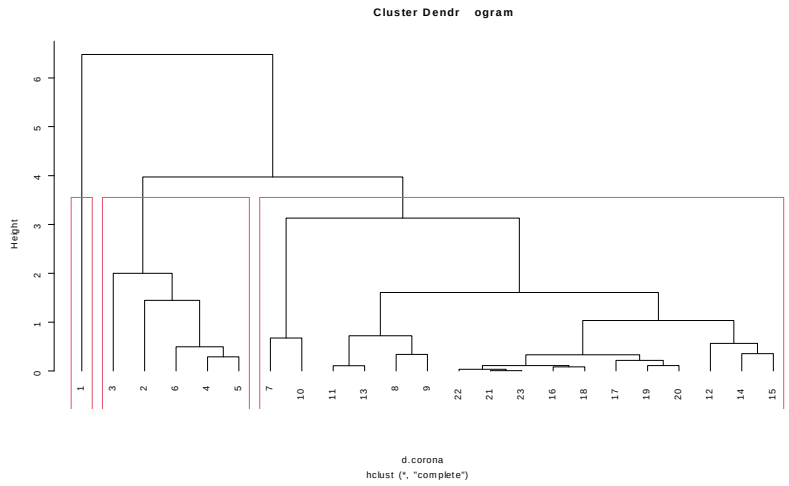

FIG. 2.7 – CAH par la méthode du saut maximal.

Commentaire :

$1^{ére}$ **classe :** L'Irak, le pays le plus contaminé et dont le nombre de mortalité est très élevé.

$2^{ème}$ **classe :** Elle concerne les pays où leurs nombre de contamination est assez faible par rapport à la première classe, mais elle reste quand même supérieure en la comparant avec la troisième classe.

$3^{ème}$ **classe :** Elle concerne les pays les moins contaminés et par conséquent, leurs nombre de guérison est assez élevé.

Les résultats obtenus par cette méthode sont un peu différents à ceux des deux méthodes précédentes.

Méthode du saut minimal :

```
cah.single=hclust(d.corona, method="single",members=NULL)
plot(cah.single, hang=-1)              # Dendrogramme de la méthode du saut minimal.
rect.hclust(cah.single,k=3)            # Découpage en trois sous groupes.
dcah.single=cutree(cah.single,k=3)     # Découpage numérique en trois sous groupes.
```

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1	2	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2

Graphe :

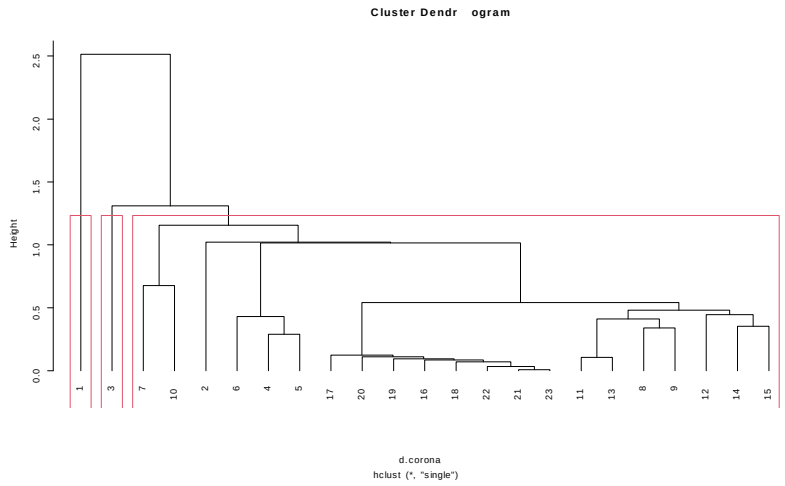

FIG. 2.8 – CAH par la méthode du saut minimal.

Commentaire :

$1^{ére}$ **classe :** L'Irak, le pays le plus contaminé et dont le nombre de mortalité est très élevé.

$2^{ème}$ **classe :** L'UAE, le pays où le nombre de contamination est élevé mais assez faible par rapport à la première classe.

$3^{ème}$ **classe :** Elle concerne les pays les moins contaminés par rapport aux deux précédentes classes.

Les résultats obtenus par cette méthode sont totalement différents à ceux des méthodes précédentes.

Classification non hiérarchique :

```
cnh.kmeans=kmeans(corona.cr,centers=3,nstart=5)  # Méthode des centres mobiles.
cnh.kmeans$betweenss      # Inertie intreclasse.
45.69342
cnh.kmeans$tot.withinss   # Inertie intraclasse.
20.30658
cnh.kmeans$totss          # Inertie totale.
66
print(cnh.kmeans)         # Découpage numérique en trois sous groupes.
```

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
2	2	1	2	2	2	2	1	1	2	1	3	1	1	3	3	3	3	3	3	3	3	3

```
inertie.exp=rep(0,times=10)  # Evaluation de la proportion d'inertie expliquée.
for (i in 2 :10)  {
                  clus=kmeans(corona.cr,centers=i,nstart=5)
                  inertie.exp[i]=clus$betweenss/clus$totss
                  }
plot(1 :10,inertie.exp,type="b",xlab="Nb. de groupes",ylab="%inertie expliquée")
```

Graphe :

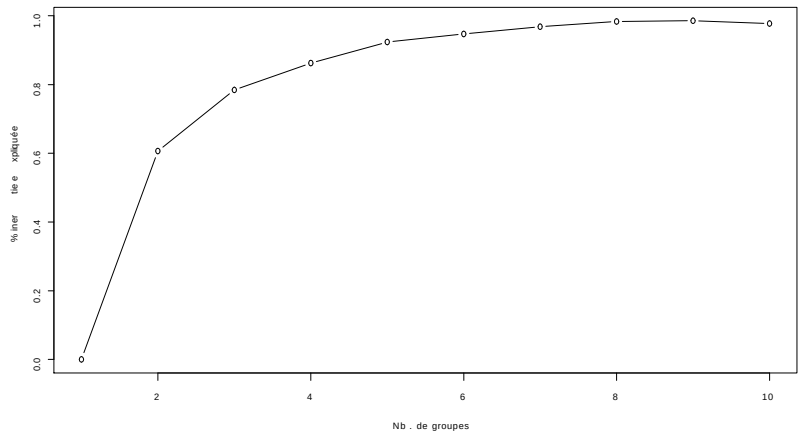

FIG. 2.9 – Représentation de l'inertie.

Commentaire

A partir de k=3 classes, l'adjonction d'un groupe supplémentaire n'augmente pas "significativement" la part d'inertie expliquée par la partition.

Comparaison entre les méthodes hiérarchiques et non hiérarchique :

print(table(dcah.ward,cnh.kmeans$cluster)) # Ward et k-means.

dcah.ward	1	2	3
1	0	1	0
2	0	0	7
3	15	0	0

print(table(dcah.ave,cnh.kmeans$cluster)) # Liaison moyenne et k-means.

dcah.ave	1	2	3
1	0	1	0
2	0	0	7
3	15	0	0

print(table(dcah.complete,cnh.kmeans$cluster)) # Saut maximal et k-means.

dcah.complete	1	2	3
1	0	1	0
2	0	0	5
3	15	0	2

print(table(dcah.single,cnh.kmeans$cluster)) # Saut minimal et k-means.

dcah.single	1	2	3
1	0	1	0
2	15	0	6
3	0	0	1

Conclusion générale :

K-means, à la différence de la CAH, ne fournit pas d'outil d'aide à la détection du nombre de classes. On doit le fixer à l'avance, en suivant différents algorithmes. La procédure est souvent la même, on fait varier le nombre de groupes et on surveille l'évolution d'un indicateur de qualité de la solutiion i.e l'aptitude des individus à être plus proches de ses congénères du même groupe que les individus des autres groupes.

On conclut que parmi toutes ces méthodes de classifications, la méthode de ward et la méthode de k-means donnent des résultats similaires et une très bonne classification des données.

Conclusion

Pour mieux comprendre le rôle de tonte ces méthodes, une implémentation a été faite sons logiciel R, sur des personnes atteintent du coronavirus.

On a conclut que :

La méthode de classification hiérarchique est facile à exécuter car elle n'exige aucune condition sur le choix du nombre de classe (qui doit être assez petit), inversement à la méthode de classification non hiérarchique au le nombre de classe doit être fixé à l'avance.

La méthode de classification non hiérarchique permet de traiter rapidement un nombre d'objets assez élève, que la classification hiérarchique devient difficile voir impossible à réaliser.

Bibliographie

[1] Celeux, G., Diday, E., Govaert, G. (1989). *Classification automatique des données environnement statistique et informatique*. Dunod, Informatique.

[2] Diday, E. (1971). *Une nouvelle méthode en classification automatique et reconnaissance des formes la méthode des nuées dynamiques*. Revue de statistique appliquée, vol. 19, No 2, p. 19-33.

[3] Dice, L., R. (1945). *Measures of the amount of ecologic association between species*. Ecology, vol. 26, No. 3, p. 297-302.

[4] Forgy, E., W. (1965). *Cluster analysis of multivariate data : effciency versus interpretability of classifications*. Biometrics, vol. 21, p. 768-769.

[5] Gersho, A., Gray, R., M. (1992). Constrained Vector Quantization. Springer, Boston, MA, p. 407-485.

[6] Godfrey, N., L., William, W., T. (1966). *Computer programs for hierarchical polythetic classification (similarity analysis)*. Computer Journal, vol. 9, No. 1, p. 60-64.

[7] Godfrey, N., L., William, W., T. (1967). *Mixed-data classificatory programs |- Agglomerative Systems*. Australian Computer Journal, vol. 1, no. 1, p. 15-20.

[8] https ://www.sis.gov.eg.

[9] Jaccard, P. (1908). *Nouvellesrecherchessur la floraledistribuição*. Bul. Soc. Science VaudoiseNaturales, vol. 44, p. 223-270.

[10] Mac-Queen, J., B. (1967). *Some Methods for classification and Analysis of Multivariate Observations*. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, No. 14, p. 281-297.

[11] Mahalanobis, P., C. (1936). *On the generalised distance in statistics*, Proceedings. of the National Institute of Sciences of India, Vol. 2, p. 49-55.

[12] Minkowski, H. (1896). *Sur les propriétés des nombres entiers qui sont dérivées de l'intuition de l'espace. Nouvelles annales de mathématiques*. journal des candidats aux écoles polytechnique et normale, Vol. 15, p. 393-403.

[13] Russel, P., F., RAO, T., R. (1940). *On habitat and association of speciesof anophelinae larvae in south-eastern Madras*. J. Malaria Inst. India, Vol. 3, No. 1, p. 153-178.

[14] Sneath, P., H., Sokal, R., R. (1973). *The principles and practice of numerical classification*. Numerical taxonomy.

[15] Ward, J., H. (1963). *Hierarchical grouping to optimize an objective function.* Journal of the American statistical association, Vol. 58, No. 301, p. 236-244.

[16] Williams, W., T., Lambert, J., M. (1959). *Multivariate methods in plant ecology : I. Association-analysis in plant communities*. The Journal of Ecology, p. 83-101.

Annexe A : Logiciel *R*

R est un système, communément appelé langage et logiciel, qui permet de réaliser des analyses statistiques. Plus particulièrement, il comporte des moyens qui rendent possible la manipulation des données, les calculs et les représentations graphiques. R a aussi la possibilité d'exécuter des programmes stockés dans des fichiers textes et comporte un grand nombre de procédures statistiques appelées paquets. Ces derniers permettent de traiter assez rapidement des sujets aussi variés que les modèles linéaires (simples et généralisés), la régression (linéaire et non linéaire), les séries chronologiques, les tests paramétriques et non paramétriques classiques, les différentes méthodes d'analyse des données,... Plusieurs paquets, tels *ade4, FactoMineR, MASS, multivariate, scatterplot3d et rgl* entre autres sont destinés à l'analyse des données statistiques multidimensionnelles.

Annexe B : Tableau des données

Le tableau ci-dessous représente les données sur le coronavirus appelé aussi Covid-19 de 23 pays arabes selon le nombre de personnes contaminées, rétablies et décédées.

	Pays	Blessures	Mortalité	Guérison
1	Irak	1.201.352	16.375	1.116.456
2	Jordanie	736.534	9.462	715.796
3	UAE	570.836	1.68	550.525
4	Liban	540.277	7.723	517.859
5	Maroc	519.108	9.143	506.962
6	Arabie saoudite	450.436	7.362	433.413
7	Tunisie	344.688	12.623	301.869
8	Palestine	308.048	3.495	300.661
9	Koweït	307.812	1.771	292.701
10	Égypte	261.666	15.047	191.475
11	Qatar	217.458	556	213.336
12	Oman	217.224	2.345	199.96
13	Bahreïn	238.156	953	208.445
14	Libye	185.776	3.126	171.874
15	Algérie	128.725	3.465	89.625
16	Soudan	35.495	2.63	29.27
17	Syrie	24.467	1.766	21.598
18	Mauritanie	19.494	463	18.475
19	Somalie	14.66	769	6.764
20	Yémen	6.742	1.321	3.445
21	Soudan du Sud	10.688	115	10.514
22	Comores	3.881	146	3.719
23	Djibouti	11.528	154	11.364

TAB. 2.1 – Statistiques sur le coronavirus dans les pays arabes.

Annexe C : Notation et obréviation

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous.

Symbole	Signification
Ω	Ensemble des individus.
$\emptyset$	Ensemble vide.
$\mathbb{R}$	Ensemble réel.
X	Tableau des données.
k	Nombre de variable.
n	Nombre d'individu.
i.e	C'est-à-dire.
e_i	Vecteur des individus.
x_j	Vecteur des variables.
X_c	Matrice des données centrées.
X_r	Matrice des données centrées réduites.
$\bar{x}$	Moyenne arithmétique.
p	Matrice des poids.
I_n	Matrice indicatrice d'ordre n.
Cov	Covariance.

Var	Variance.
σ	Ecart type.
V	Matrice de variance-covariance.
V^{-1}	Inverse de la matrice de variance-covariance.
D_σ	Matrice diagonale de la racine de la variance.
D_σ^{-1}	Inverse de matrice diagonale de la racine de la variance.
Cor	Corrélation.
R	Matrice de corrélation.
d	Distance.
$\alpha, \beta, \gamma, \delta$	Differentes caracteristiques.
η	Nombre totale des caracteristiques.
$\mathbf{H}$	Hiérarchie.
D	Matrice des distances.
C_i	Classe i.
n_{C_j}	Nombre d'objet du groupe C_j pour $j = 1, 2$.
g	Centre de gravité.
q_k	Poids de C_j.
g_j	Centre de gravité de C_j.
I_{tot}	Inertie totale.
I_{inter}	Inertie inter-classe.
I_{intra}	Inertie intra-classe.
P_k	Partition.
CAH	Classification ascendante hiérarchique.
CDH	Classification descendante hiérarchique.

Résumé

La méthode de classification est une méthode statistique exploratoire qui permet de regrouper un ensemble d'individus décrits par des caractéristiques. Il existe deux principales méthodes de classification: classification hiérarchique et non hiérarchique. La classification hiérarchique peut être ascendante ou descendante, elle consiste à effectuer une suite de regroupement en classes de moins en moins fines en agrégeant à chaque étape les objets ou les groupes d'objets les plus proches. Le nombre d'objets n'est pas fixé au préalable. Tandis que la classification non hiérarchique, exige que le nombre de groupes doit être fixé dès le départ. Elle se distingue par une multitude d'algorithmes.

Mots clés : Classification; Hiérarchie; Similarité; Dissimilarité; Distance euclidienne; Dendrogramme; Objet; Centre de gravité; Centre mobile.

Abstract

The classification method is an exploratory statistical method that allows a grouping of individuals set described by characteristics. There are two main methods of classification: hierarchical and not hierarchical. The hierarchical classification can be ascending or descending, it consists in carrying out a series of groupings into less and less fine classes by aggregating at each step the closest objects or groups of objects. The number of objects is not fixed beforehand. Whereas non-hierarchical classification requires that the number of groups must be fixed from the start. She stands out by a multitude of algorithms.

Key words: Classification; Hierarchy; Similarity; Dissimilarity; Euclidean distance; Dendrogram; Object; Gravity center; Mobil center.

ملخص

طريقة التصنيف هي طريقة إحصائية استكشافية تسمح بتجميع مجموعة من العناصر ذات صفات و خصائص معينة.

هناك طريقتان رئيسيتان للتصنيف: تصنيف هرمي وغير هرمي. التصنيف الهرمي يمكن أن يكون تصاعديًا أو تنازليًا، ويتألف من القيام بسلسلة من التجمعات في فئات أقل وأقل دقة من خلال تجميع أقرب الأجسام أو مجموعات الأجسام في كل خطوة من دون تحديد مسبق لعدد العناصر. في حين التصنيف غير الهرمي يتطلب تحديد عدد المجموعات من البداية. ويتميز هذا الأخير بعدد كبير من الخوارزميات.

الكلمات المفتاحية: التصنيف التسلسل الهرمي، التشابه، الإختلاف، المسافة الإقليدية، مخطط شجري، العنصر، مركز الثقل، مركز متنقل