11.

$$CTR_1(i_1) = \frac{1}{3} \times \frac{3/4}{1/2} = 1/2 \textit{ et } CTR_2(i_1) = \frac{1}{3} \times \frac{3/36}{1/6} = 1/6 \tag{2.1}$$

Comme il n'y a que deux axes non triviaux , la qualite de représentation de i_1 dans le plan factoriel 1-2 est 1.

Chapitre 4

Classification

4.1 Introduction

Ce chapitre concerne les méthodes de classification ou *cluster analysis* en anglais. Le but est de regrouper les individus en classes qui sont le plus homogène possible.
Les méthodes de classification font parties intégrante de l'analyse de données est depuis longtemps forme une problématique importante issue surtout de l'étude des phénomènes naturelles . Les techniques de classification sont des méthodes permettant de regrouper les lignes ou les colonnes d'un tableau de données sur base d'une proximité entre ces lignes ou ces colonnes.
Selon la nature du tableau de données, la distance utilisée pour juger de la proximité de nos éléments statistiques sera différente. Par exemple, si les données sont toutes de nature quantitative, une distance de type euclidienne pourra être utilisée ; pour des données nominales par contre, nous devrons faire appel à des indices de distance basés sur la construction de tableaux croisés. par exemple si les variables sont de type échelle (quantitative), une distance euclidienne standard peut s'appliquer. Si les variables sont nominales, on utiliser une distance de type Khi-deux.
D'une manière générale, les techniques de classification sont donc des méthodes mathématiques qui cherchent à effectuer des regroupements des individus statistiques les plus proches dans un espace à dimensions multiples.

La classification a pour principal objectif de regrouper les individus décrits par un ensemble de variables, ou regrouper les variables observées sur des individus et d'interpréter ces regroupements par une synthèse des résultats. L'intérêt de regrouper les individus

est ici de les classer en conservant leur caractère multidimensionnel, et non pas seulement à partir d'une seule variable. Si les variables sont nombreuses il peut être intéressant de les regrouper afin de réduire leur nombre pour une interprétation plus facile.
Les méthodes de classification sont donc complémentaires des analyses factorielles décrites dans les chapitres précédents.

4.1.1 Domaines d'application

Ces techniques peuvent être utilisées dans de nombreux domaines comme :

— le domaine médicale : regrouper des patients afin de définir une thérapeutique adaptée à un type particulier de malades

— le domaine du marketing, afin de définir un groupe cible d'individus pour une campagne publicitaire. On parle alors souvent de segmentation à ne pas confondre avec la segmentation qui est une méthode de discrimination par arbre.

— Dans le domaine de la reconnaissance des formes elle porte le nom de *classification non-supervisée*.

En biologie on parle souvent de *Taxonomie* et en Intelligence Artificielle on parle *d'Apprentissage non supervisé*. Il ne faut pas confondre les méthodes de classification avec les méthodes explicatives de discrimination dont l'objectif est d'expliquer une partition connue à priori, c'est à dire d'expliquer une variable qualitative dont chaque modalité décrit une classe de la partition, par un ensemble de variables de type quelconque (et non pas d'expliquer une variable qualitative comme en régression).

En anglais, le terme désignant les méthodes de classification automatique est *Clustering* les classes étant des *clusters*.
Ce terme En anglais désignant les méthodes de discrimination est Classification et en Intelligence Artificielle on parle d'Apprentissage supervisé ou encore de Reconnaissance des formes (Pattern Recognition). Aujourd'hui le Data Mining (synonymes : Fouille de données, extraction de connaissance, ...) et un champs d'application à l'interface de la statistique et des technologies de l'information (bases de données, Intelligence Artificielle, apprentissage). On définie parfois le Data Mining comme l'extraction de connaissances de grandes bases de données. Le Data Mining utilise donc souvent les méthodes d'Analyse de données comme les méthodes de classification, l'analyse discriminante.

4.1.2 Les données

Les données de départ sont souvent organisées comme une matrice X de dimension (n$\times$ p) (n individus $\times$ p variables), où x_{ij} est la valeur de la variable j pour l'individu i,

$$\begin{array}{c} \\ 1 \\ \vdots \\ i \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{cccc} 1 & \dots & j & \dots & p \end{array} \\ \begin{pmatrix} & & & \\ & & \vdots & \\ \dots & x_{ij} & \dots & \\ & \vdots & & \\ & & & \end{pmatrix} \end{array}$$

Les variables peuvent être quantitatives continues ou issues de tableaux de contingences, ou binaires issues de tableaux logiques, ou encore qualitatives. Afin de traiter l'ensemble de ces types de variables, c'est la mesure de similarité ou dissimilarité qui doit être adaptée aux types de données. Une mesure de similarité ou de dissimilarité est une distance à l'exception que l'inégalité triangulaire n'est pas exigée. Ces mesures peuvent être des distances dans le cas de variables quantitatives. Ainsi,

Il est préférable d'employer une distance euclidienne, de Mahalanobis où de Minkowsky pour les variables quantitatives continues et une distance du χ^2 pour des tableaux de contingences, distances que nous avons déjà présentées dans chapitre précédent.

Si le tableau est composé de données mixtes, il suffit de rendre les variables quantitatives en variables qualitatives en choisissant quelques modalités et pour se faire, il suffit de découper l'intervalle de variation en sous-intervalles qui définissent autant de modalités. Ainsi diminuer le nombre de classes, c'est regrouper des individus de plus en plus différents et augmenter le nombre de classes, c'est obtenir des classes plus nombreuses et à faible effectif.

On considère un ensemble $\Omega = \{1, ..., i, ...n\}$ de n individus décrits par p variables $X^1, ..., X^p$ dans une matrice X de n lignes et p colonnes .

Définition 4.1.1. *. Il y a deux grands types de méthodes de classification :*

— ***Classifications hiérarchiques.*** *A chaque instant, on a une décomposition de l'espace des individus en classes disjointes. Ces méthodes peuvent être ascendantes ou descendantes. Dans ce cas, au début, chaque individu forme une classe à lui tout seul. Puis, à chaque étape, les deux classes les plus "proches" sont fusionnées. A la dernière étape, il ne reste plus qu'une seule classe regroupant tous les individus.*

— ***Classifications non-hiérarchiques.*** *Dans ce cas, le nombre de classe est fixé à l'avance. Il s'agit essentiellement des techniques d'agrégation autour des centres mobiles. Ces méthodes sont particulièrement intéressantes dans le cas de grands tableaux car elles sont peu coûteuses en temps de calcul et en espace de mémoire.*

Dans ces méthodes, les individus sont regroupés dans des classes homogènes. Ceci signifie que les individus d'une même classe sont proches. Afin de mesurer la proximité entre individus, nous avons besoin d'un choix de distance, objet de la section suivante.

.

4.1.3 Distances

Voici les distances les plus utilisées. Soit x_i, x_i' une paire d'individus.

1. La distance usuelle où bien la distance euclidienne simple :
   Les distances définies par : $d^2(x_i, x_i') = \| x_i - x_i' \|_M^2 = (x_i - x_i')^t M (x_i - x_i')$
   — Si $M = I_p$: d est la distance euclidienne simple,
   — si $M = D_{1/\sigma^2}$ (matrice diagonale des inverses des variances empiriques des p variables), on se ramène à une distance euclidienne simple entre variables réduites (j ième colonne divisée par l'écart-type empirique σ_j). On parle de distance euclidienne normalisée par l'inverse de la variance
   — si $M = V^{-1}$, d est la distance de Mahalanobis.

   $$d(x_i, x_{i'}) = \| x_i - x_{i'} \|_{V^{-1}},$$

   où V est la matrice des variances.
2. La distance de city-block ou de Manhattan : $d(x_i, x_{i'}) = \sum_{j=1}^{p} |x_{ij} - x_{i'j}|$.
3. La distance de Chebychev, ou distance du $\max$: $d(x_i, x_{i'}) = \max_{j=1\ldots p} |x_{ij} - x_{i'j}|$.

Remarque 4.1.1. — *En général, on utilise la distance euclidienne lorsque tous les paramètres ont une variance équivalente. En effet, si une variable a une variance bien plus forte, la distance euclidienne simple va accorder beaucoup plus d'importance à la différence entre les deux individus sur cette variable qu'à la différence entre les deux individus sur les autres variables. Il est préférable dans ce cas d'utiliser la distance euclidienne normalisée par l'inverse de la variance, afin de donner la même importance à toutes les variables. Cela revient à réduire tous les variables (les diviser par leur écart-type) et à calculer ensuite la distance euclidienne simple.*

— *Si les unités des variables ne sont pas comparables, on pourra aussi considérer les données centrées-réduites. Une autre alternative est de prendre le tableau issu de l'ACP.*

— *Par définition, une distance satisfait aux conditions suivantes :*

1. *Symétrie :* $d(x_i, x_{i'}) = d(x_{i'}, x_i)$.
2. *Positivité :* $d(x_i, x_{i'}) \geq 0$.
3. $d(x_i, x_{i'}) = 0 \Longleftrightarrow x_i = x_{i'}$.
4. *Inégalité triangulaire :* $d(x_i, x_{i'}) = d(x_i, x_k) + d(x_k, x_{i'})$.

Dans certain cas, il est utile de relaxer la dernière hypothèse, on parlera alors de dissimilarité. Par exemple, une question importante en biologie est celle de la classification des espèces. Les n individus sont alors décrits par la présence (1) ou absence (0) de p caractéristiques. Il sŠagit dŠun tableau disjonctif complet. Les distances ci-dessus ne sont pas adéquates,

4.2 La classification hiérarchique (CAH)

La classification hiérarchique ascendante (Hierarchical Cluster Analysis) est une méthode itérative qui consiste, à chaque étape, à regrouper les classes les plus proches. A la première étape chaque individu constitue une classe. L'algorithme s'arrête avec l'obtention d'une seule classe. Les regroupement successifs sont représentés sous la forme d'un arbre ou dendogramme.

4.2.1 Le principe

La classification hiérarchique procède à un regroupement d'individus caractérisés par des critères des variables.

Lors de la première étape, chaque individu est considéré comme une classe à part entière. Nous avons donc, à ce niveau du processus, autant de classes que d'individus $\Longrightarrow$ ***N classes pour N individus***

L'algorithme de classification hiérarchique commence par calculer *une distance entre toutes les classes (ou tous les individus)* présentes dans le tableau de données. Généralement, la distance euclidienne est utilisée afin de rendre compte de cette mesure. Ainsi, **plus cette distance sera petite, plus les classes seront proches**, et, donc, seront similaires.

Une fois l'ensemble des distances entre les points calculées, l'algorithme *va fusionner les deux individus (ou les deux classes)* ayant *la distance la plus petite* (donc les plus semblables) pour ne constituer qu'une seule classe. Ainsi, à la fin de cette première étape, une classe a disparu $\Longrightarrow$ ***N-1 classes pour N individus.***

L'algorithme repart à zéro puisqu'il recalcule, à nouveau, toutes les distances entre les classes, pour fusionner deux nouvelles classes, selon le même principe que précédemment (ie les classes dont les distances sont les plus petites). A la fin de cette deuxième étape, nous avons $\Longrightarrow$ ***N-2 classes pour N individus***.

Ce processus continue jusqu'à ce qu'il ne reste plus qu'une seule classe. En d'autres termes, toutes les classes finissent, en fin d'algorithme par ne constituer qu'une seule classe $\Longrightarrow$ ***1 classe pour N individus.***

Tout le travail de l'analyste constituera, maintenant, à remonter les étapes à partir de la classe 1 pour savoir à quel niveau il est pertinent de s'arrêter. Nous voyons bien en quoi les individus sont classés d'une façon hiérarchique : au début, les individus les plus proches sont regroupés. Plus nous avançons dans les étapes, plus les individus diffèrent. L'interprétation du nombre de classes se fait donc toujours à partir des dernières étapes.

Exemple 4.2.1. *Soit Ω un ensemble avec 9 éléments : a, b, ..., i.*

Supposons que la CHA a produit la suite de partitions suivante :

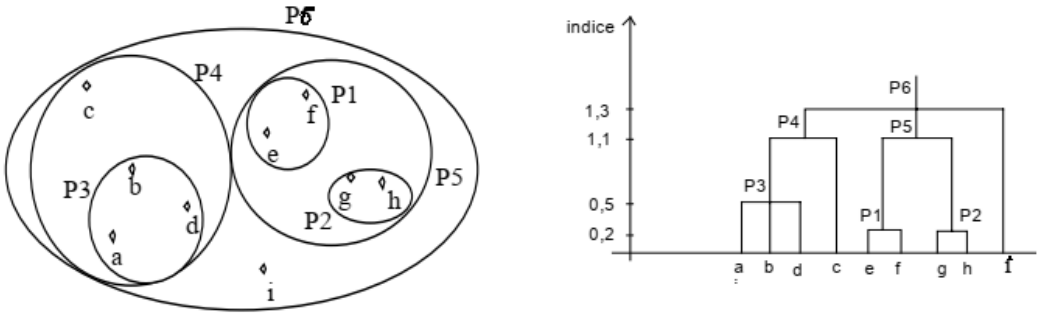
niveau 0 *a b c d e f g h i*

niveau 1 *a b c d (e f) (g h) i*

niveau 2 *(a b d) c (e f) (g h) i*

niveau 3 *(a b d c) (e f g h) i*

niveau 4 *(a b c d e f g h i)*



Le résultat de l'algorithme est représenté sous forme d'un *arbre*, aussi appelé ***dendogramme***. La hauteur des branches représente la distance entre les deux éléments regroupés.

Les avantages et inconvénients de l'algorithme sont :

— *Avantages.* Algorithme simple, permettant une très bonne lecture des données.

— *Inconvénients.*

1. Selon la définition de distance entre les classes, on trouve des résultats très différents. Une idée est donc d'appliquer la méthode avec différentes distances, et de trouver les groupes stables.
2. L'algorithme est constitué de n étapes, et à chaque fois il faut parcourir la matrice des distances qui est de taille $\frac{(n-k)(n-k-1)}{2}$

4.2.2 Le choix de la métrique

Le choix de la métrique dépend entièrement du type de variables que l'algorithme de classification hiérarchique accepte. Comme il accepte 3 types différents de variables (continues, nominales ou dichotomiques), nous aurons 3 familles de mesure, appelés respectivement Intervalle, Effectifs et Binaire.

— Pour les variables continues, on utilise traditionnellement *une distance euclidienne* car facile à comprendre et facile à calculer.

Si deux individus ont 35 et 45 ans et ont débuté à travailler à 21 et 22 ans, la distance euclidienne entre ces deux individus sera égale à $d(x,y) = \sqrt{(35-45)^2 + (21-22)^2}$

soit $10,05$.

Il existe un grand nombre de distances plus où moins utilisées. Pour les variables continues, nous pouvons citer le coefficient de corrélation de Pearson, Cosinus, Distance de Tchebycheff, etc...

— Pour les variables nominales, de nouveau, il est souvent plus simple de se rapporter à des choses connues comme, par exemple, une distance du χ^2.

— Enfin, pour les variables du type dichotomique, le choix, par défaut, se porte sur une distance euclidienne.

4.2.3 Choix de la méthode

On a la question sur quel critère repose le regroupement des individus en classes ? Lors du calcul des distances entre deux classes qui comprennent plusieurs individus, il existe plusieurs possibilités en matière de choix de points de référence de la classes .

Définition 4.2.1. *On appelle* stratégied'agrégation*, la façon d'apprécier la proximité entre deux classes C_1 et C_2, au cours des agrégations successives qui se réalisent lors de la construction de la hiérarchie.*

Voici quelques définitions naturelles de distances entre classes. Soient C et C' deux classes.

— *La distance du saut minimum* (Single Linkage ou Nearest Neighbor) calcule les distances entre les points pour regrouper les classes dont les distances entre les points sont les plus petites. L'inconvénient de ce type de méthode est son incapacité de différencier des classes proches.

la distance entre les classes C et C', notée $d_{\min}(C, C')$, est par définition :

$$d_{\min}(C, C') = \min_{x_i \in C, x_{i'} \in C'} d(x_i, x_{i'})$$

C'est la plus petite distance entre éléments des deux classes.

— *La distance du saut maximum* où bien La méthode suivant le diamètre (Complete Linkage ou Furthest Neighbor) prend la démarche inverse, c'est-à-dire qu'une fu-

sion entre deux classes s'opère lorsque les distances entre deux points de deux classes différentes sont les plus éloignées. Comme la technique précédente, cette méthode est relativement insensible aux valeurs extrêmes.
la distance entre les classes C et C', notée $d_{\max}(C, C')$, est par définition :

$$d_{\max}(C, C') = \max_{x_i \in C, x_{i'} \in C'} d(x_i, x_{i'})$$

C'est la plus grande distance entre éléments des deux classes.

%item Une autre méthode qui peut paraître plus intuitive est celle des barycentres (centroïds)

— La méthode par défaut, *la distance moyenne entre les classes (Between-groups Average Linkage ou Baverage)* possède également des propriétés intéressantes puisqu'elle gère relativement bien les bruits. Par contre, son inconvénient est qu'elle est influencée par les valeurs extrêmes. Son principe est de prendre en compte une moyenne de distances entre les classes, pour chaque individu.

— La dernière méthode dont nous pouvons dire quelques mots est *la méthode de Ward*. Ses propriétés sont assez proches de la méthode des distances moyennes entre les classes. Son approche est intéressante puisqu'elle repose sur la décomposition de la variance. En effet, une variance comporte deux éléments : une partie qui explique les différences entre les classes (appelée variance inter classe ou expliquée) et une autre qui relate les différences dans les groupes (variance intra classe ou résiduelle).

On suppose que les individus sont attribués des poids relatifs (ω_i), avec $\sum_{i=1}^{n} \omega_i = 1$, et que la distance entre individus est la distance euclidienne usuelle.
L'inertie totale du nuage de points est :

$$I_{tot} = \sum_{i=1}^{n} \omega_i d^2(xi, g) = \sum_{i=1}^{n} \omega_i \parallel x_i - g \parallel^2 .$$

Supposons de plus que les individus soient regroupés en k classes $C_1, ..., C_k$. Pour tout $\ell \in \{1, ..., k\}$:

— $\omega_{(\ell)}$ est le poids relatif de la classe C_ℓ : $\omega_{(\ell)} = \sum_{i \in I_\ell} \omega_i$ ainsi , $\sum_{\ell=1}^{k} \omega_\ell = 1$

— $g_{(\ell)}$ est le centre de gravité ou barycentre de la classe $C_{(\ell)}$ où pour tout $j \in \{1, ..., p\}$, :

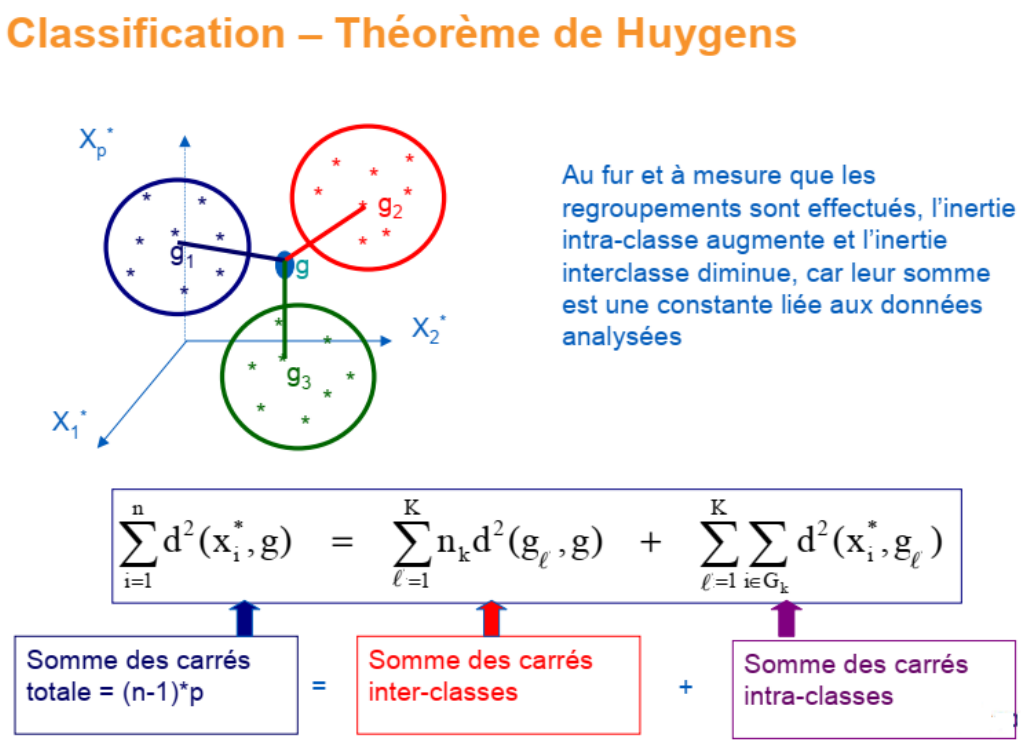
$$g_{(\ell)} = \frac{1}{\omega_{(\ell)}} \sum_{i \in I_\ell} \omega_i x_{ij}$$

L'inertie inter-classes du nuage de points est :

$$I_{inter} = \sum_{\ell=1}^{k} \omega_{(\ell)} d^2(g_{(\ell)}, g) = \sum_{\ell=1}^{k} \omega_{(\ell)} \parallel g_{(\ell)} - g \parallel^2 .$$

Nous avions aussi vu la notion d'inertie intra-classes et le théorème de Huygens, qui dit que :

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



Exemple 4.2.2. *Supposons qu'on ait 5 individus et une matrice des dissimilarités donnée par*

$$D = \left(\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & 0 & 7 & 6 & 3 & 2 \\ 2 & 7 & 0 & 1 & 10 & 5 \\ 3 & 6 & 1 & 0 & 9 & 4 \\ 4 & 3 & 10 & 9 & 0 & 5 \\ 5 & 2 & 5 & 4 & 5 & 0 \end{array}\right)$$

D'abord, le saut minimum sera utilisé.

Étape 1 : *(cette étape est indépendante du type de distance entre les classes employée !) chaque individu est dans une classe (i.e. $C_i = \{i\}$), la matrice des distance est ainsi D. La* plus petite distance est ***1****. Elle correspond aux deux éléments à plus petite distance : 2 et 3. Ils sont alors fusionnés, et les classes sont maintenant* $\{1\}, \{2,3\}, \{4\} et \{5\}$.

Étape 2 : *Une nouvelle matrice des distances doit être calculée (selon la règle du saut minimum).*

$\backslash$	$\{1\}$	$\{2,3\}$	$\{4\}$	$\{5\}$
$\{1\}$	0	6	3	***2***
$\{2,3\}$	6	0	5	4
$\{4\}$	3	5	0	5
$\{5\}$	2	4	5	0

La plus petite entrée est 2, et elle correspond aux classes {1} et {5}. Ces classes seront fusionnées pour donner les nouvelles classes $\{1,5\}$, $\{2,3\}$ *et* $\{4\}$.

$\backslash$	$\{1,5\}$	$\{2,3\}$	$\{4\}$
$\{1,5\}$	0	4	***3***
$\{2,3\}$	4	0	5
$\{4\}$	***3***	5	0

Sa plus petite entrée est le 3, et en conséquence les classes de {1,5} et {4} seront fusionnées. On a maintenant deux classes : {1,4,5} et {2,3}.

Étape 4 : *Il ne reste que deux classes, et donc une seule possibilité pour faire la fusion. Néanmoins, il est utile de connaître la distance qui les sépare : ici 4.*

Exemple 4.2.3. *Le même exemple avec le saut maximum :*

Étape 1 : *Comme à l'exemple précédent : la plus petite distance (celle entre 2 et 3) est 1. Ils sont alors fusionnés, et les classes sont maintenant {1},{2,3},{4} et {5}.*
Étape 2 : *Une nouvelle matrice des distances doit être calculée (selon la règle du saut maximum).*

$\backslash$	$\{1\}$	$\{2,3\}$	$\{4\}$	$\{5\}$
$\{1\}$	0	7	3	2
$\{2,3\}$	7	0	9	10
$\{4\}$	3	9	0	5
$\{5\}$	2	10	5	0

La plus petite entrée est (encore) 2, et elle correspond aux classes {1} et {5}. Ces classes seront fusionnées pour donner les nouvelles classes {1,5},{2,3} et {4}.
Étape 3 : *La matrice des distance est*

$\backslash$	$\{1,5\}$	$\{2,3\}$	$\{4\}$
$\{1,5\}$	0	10	***5***
$\{2,3\}$	10	0	9
$\{4\}$	***5***	9	0

Sa plus petite entrée est le 5, et en conséquence les classes de {1,5} et {4} seront fusionnées. On a maintenant deux classes : {1,4,5} et {2,3}.
Étape 4 : *Il ne reste que deux classes, et donc une seule possibilité pour faire la fusion. Néanmoins, il est utile de connaître la distance qui les sépare : ici 10.*

4.3 La classification non hiérarchique

Il s'agit de regrouper n individus en k classes de telle sorte que les individus d'une même classe soient le plus semblables possible et que les classes soient bien séparées. Ceci suppose la définition d'un critère global mesurant la proximité des individus d'une même classe et donc la qualité d'une partition.

4.3.1 Méthode des centres mobiles

Cette méthode peut être vu comme un cas particulier de l'approche des nuées dynamiques développée par E. Diday (1989).
Cette méthode d'un formalisme très simple n'en est pas moins très efficace pour de vastes tableaux de données. Elle est de plus rapide, mais cependant pas toujours optimale.
La méthode des centres mobiles est fondée sur une méthode de partitionnement directe des individus connaissant par avance le nombre de classes attendues.

Principe de l'algorithme

Soit un ensemble I de n individus à partitionner, caractérisés par p caractères ou variables. On suppose que l'espace $\mathbb{R}^P$ supportant les n points individus est muni d'une distance appropriée notée d (souvent distance euclidienne usuelle ou distance du χ^2). On désire constituer au maximum q classes. Les étapes de l'algorithme sont illustrées par :

Étape 0 : On détermine q centres provisoires de classes (Le choix de ces centres est important pour la rapidité de la convergence, et les connaissances a priori doivent ici être mises à profit, s'il y en a. Dans le cas contraire, le plus courant, il suffit de tirer aléatoirement ces centres par un tirage sans remise des q centres.).

Les q centres :

$$\{C_1^0, ..., C_k^0, ..., C_q^0\}$$

induisent une première partition p^0 de l'ensemble des individus I en q classes :

$$\{I_1^0, ..., I_k^0, ..., I_q^0\}$$

Ainsi l'individu i appartient à la classe I_k^0 s'il est plus proche de C_k^0 que de tous les autres centres.

Étape 1 : On détermine q nouveaux centres de classes :

$$\{C_1^1, ..., C_k^1, ..., C_q^1\}$$

en prenant les centres de gravité des classes qui viennent d'être obtenues :

$$\{I_1^0, ..., I_k^0, ..., I_q^0\}$$

Ces nouveaux centres induisent une nouvelle partition p^1 de I construite selon la même règle que pour p^0.
La partition p^1 est formée des classes notées :

$$\{I_1^1, ..., I_k^1, ..., I_q^1\}$$

Étape m : On détermine q nouveaux centres de classes :

$$\{C_1^m, ..., C_k^m, ..., C_q^m\}$$

en prenant les centres de gravité des classes qui ont été obtenues lors de l'étape précédente,

$$\{I_1^{m-1}, ..., I_k^{m-1}, ..., I_q^{m-1}\}$$

Ces nouveaux centres induisent une nouvelle partition p^m de l'ensemble I formée des classes :

$$\{I_1^m, ..., I_k^m, ..., I_q^m\}$$

Le processus se stabilise nécessairement et l'algorithme s'arrête soit lorsque deux itérations successives conduisent à la même partition, soit lorsqu'un critère convenablement choisi (par exemple, la mesure de la variance intra-classes) cesse de décroître de façon sensible, soit encore parce qu'un nombre maximal d'itérations a été fixé *a priori.*
Généralement, la partition obtenue finalement dépend du choix initial des centres.

Il existes de nombreux algorithmes qui sont fondés sur un principe similaire. Les deux principaux sont *les nuées dynamiques* et les *k-means* ou *k-moyennes*. La différence pour la méthode des nuées dynamiques se situe au niveau de la réaffectation des individus à chaque classe.

Après avoir déterminé les centres de gravité, un *noyau* est déterminé pour chaque classe comme étant l'individu le plus proche du centre de gravité de chaque classe. La réaffectation se fait alors en fonction de la distance des autres individus aux noyaux de chaque classe. Ce formalisme a permis plusieurs généralisations de la méthode.

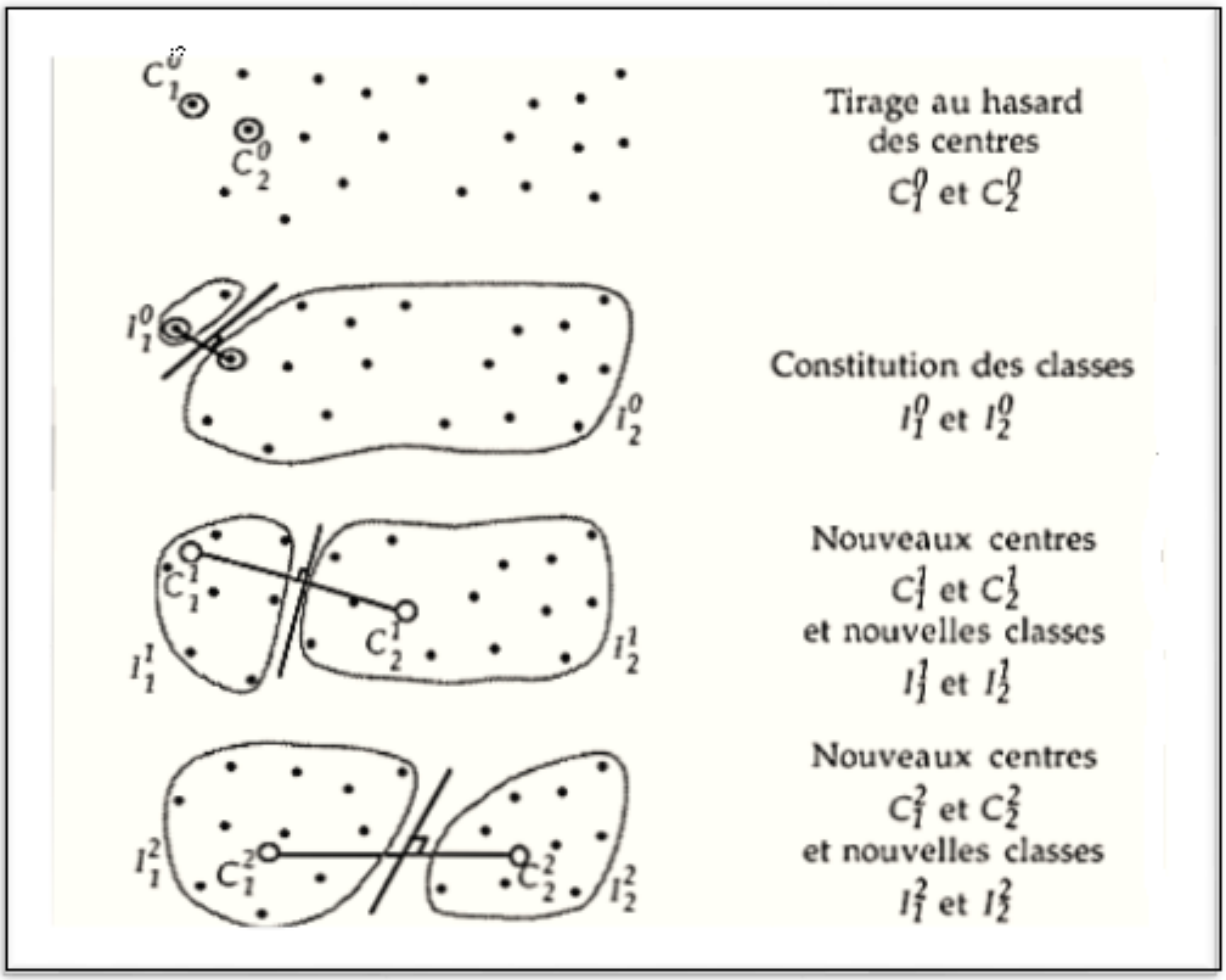


FIGURE 4.1 – Les étapes de l'algorithme

La méthode des k-means après avoir choisi une première fois les centres mobiles, recalcule le centre de chaque classe dès lors qu'un individu y est affecté. La position du centre est donc modifiée à chaque affectation, ce qui permet d'avoir une bonne partition en peu d'itérations.

L'avantage de ces méthodes est de permettre Ie traitement d'un nombre trés éleve d'elements a des coûts satisfaisants, inferieurs à ceux occasionnés pour les classifications hieérarchiques. Mais elles possèdent aussi quelques inconvénients non négligeables : à part le risque d'obtenir des classes vides, donc d'aboutir à moins de k classes, est de fournir une partition finale qui dépend de la partition de départ : on n'atteint pas l'optimum global mais seulement la meilleure partition possible à partir de celle de départ. De plus, la partition initiale est souvent arbitraire car il est courant de choisir les centres par tirage

au sort de k individus parmi n.

4.4 Interprétation

L'interprétation repose essentiellement sur la lecture du dendrogramme. Elle devient problématique lorsque le nombre d'individus est très important. Elle doit se faire de haut en bas afin d'examiner d'abord les partitions qui possèdent peu de classes, pour ensuite entrer dans des considérations plus détaillées. Nous cherchons, essentiellement la partition qui présente le plus d'intérêt. Pour cela, il faut chercher à construire des classes homogènes.

4.5 Conclusion

Nous avons dans ce chapitre présenté uniquement deux méthodes (ou famille de méthodes) de classification : la méthode des centres mobiles et la classification hiérarchique ascendante. Les méthodes de classification sont cependant très nombreuses. Il existe entre autre une méthode dite de classification mixte (*hybrid classification*) qui est un mélange de la méthode des centres mobiles et de la classification hiérarchique. Elle est particulièrement bien adaptée aux tableaux de données comportant des milliers d'individus, pour lesquels le dendrogramme est dificile à lire.

La classification est une phase importante de l'analyse des données. Il est préférable de l'employer en complément des méthodes d'analyse factorielles (particulièrement la classification ascendante hiérarchique qui utilise la méthode de Ward pour l'agrégation). Il est conseillé d'appliquer la classification après les analyses factorielles. Cependant, les classes peuvent constituer des variables supplémentaires dans l'ACP, l'AFC ou encore l'ACM.

4.6 Excercices

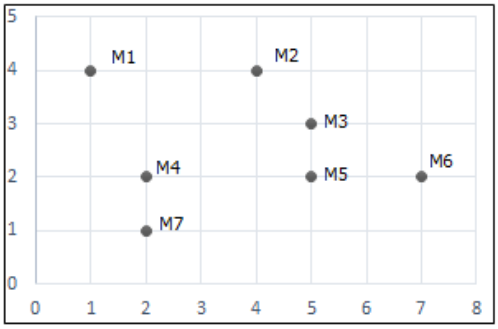
Exercice 4.6.1. *On se propose de réaliser une classification des* 7 *points suivants en utilisant la méthode d'aglomération au plus proche voisin :* $M_1 = (1;4)$, $M_2 = (4;4)$, $M_3 = (5;3), M_4 =$

$(2;2), M_5 = (5;2), M_6 = (7;2)$ *et* $M_7 = (2;1)$.

1. *Calculer le carré de la distance euclidienne de* M_1 *à* M_4.

$$d^2(M_1, M_4) = (x_1 - x_4)^2 + (y_1 - y_4)^2 = (1-2)^2 + (4-2)^2 = 5.$$

2. *Représenter graphiquement les 07 points. Calculer le tableau des distances entre les 07 points en utilisant le carré de la distance euclidienne.*



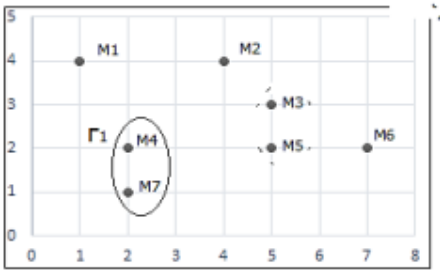
	M1	M2	M3	M4	M5	M6	M7
M1	0	9	17	5	20	40	10
M2		0	2	8	5	13	13
M3			0	10	**1**	5	13
M4				0	9	25	**1**
M5					0	4	10
M6						0	26
M7							0

3. *Sur un second dessin, en les entourant d'une courbe, les deux points les plus proches pour former une classe,* Γ_1*, puis déterminer le deuxième tableau de distance en calculant notamment les distances (au plus proche voisin) de la nouvelle classe avec les 5 autres points.*
4. *Tracer un dendrogramme résumant cette classification.*

Exercice 4.6.2. *Classification par la méthode des centres mobiles.*

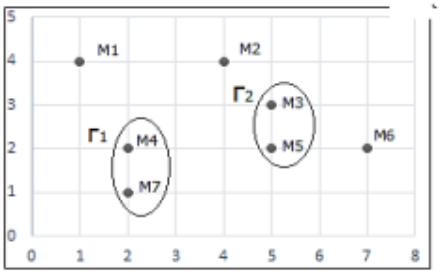
On considère les 6 points $M_1 = (2,3), M_2 = (-2,1), M_3 = (2,1), M_4 = (-1,0), M_5 = (-2,-1)$ *et* $M_6 = (2,-1)$.

1. *En supposant que les deux points* M_4 *et* M_5 *sont les centres initiaux, décrire par une succession de dessins, les étapes de l'algorithme des centres mobiles en représentant à chaque itération de l'algorithme les centres ainsi que les classes qu'on entourera chacune d'un rond.*
2. *Les trois dessins de la question précédente représentent trois partitions différentes du même ensemble. Calculer l'inertie totale du nuage puis, pour chacune des partitions, l'inertie intra classe et vérifier qu'elle est bien décroissante au cours du processus. En calculant l'inertie inter de l'une des partition, vérifier le théorème de Huygens.*



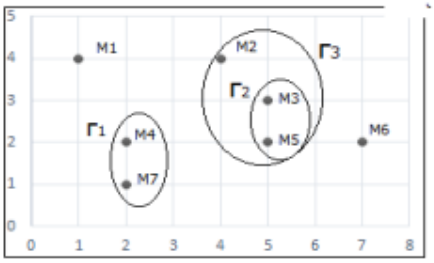
D'aprés le tableau on a deux choix (M4,M7) où (M3,M5) en prend Γ1 =(M4,M

	M1	M2	M3	M5	M6	Γ1
M1	0	9	17	20	40	5
M2		0	2	5	13	8
M3			0	(1)	5	10
M5				0	4	9
M6					0	25
Γ1						0



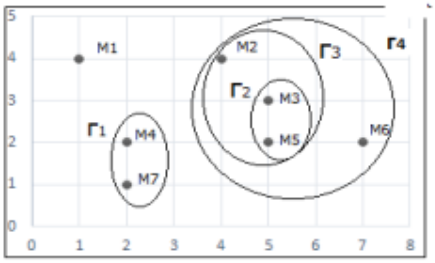
En regroupe M3 et M5 = Γ2

	M1	M2	M6	Γ1	Γ2
M1	0	9	40	5	17
M2		0	13	8	2
M6			0	25	4
Γ1				0	9
Γ2					0



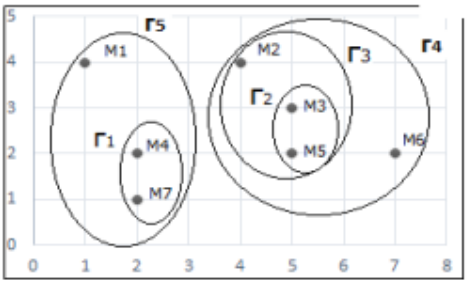
En regroupe M2 et Γ2 —— Γ3 ={M2,M3,M5}

	M1	M6	Γ1	Γ3
M1	0	40	5	9
M6		0	25	(4)
Γ1			0	8
Γ3				1



En regroupe M6 et Γ3 —— Γ4 ={M2,M3,M5 ,M6}

	M1	Γ1	Γ4
M1	0	(5)	9
Γ1		0	8
Γ4			1

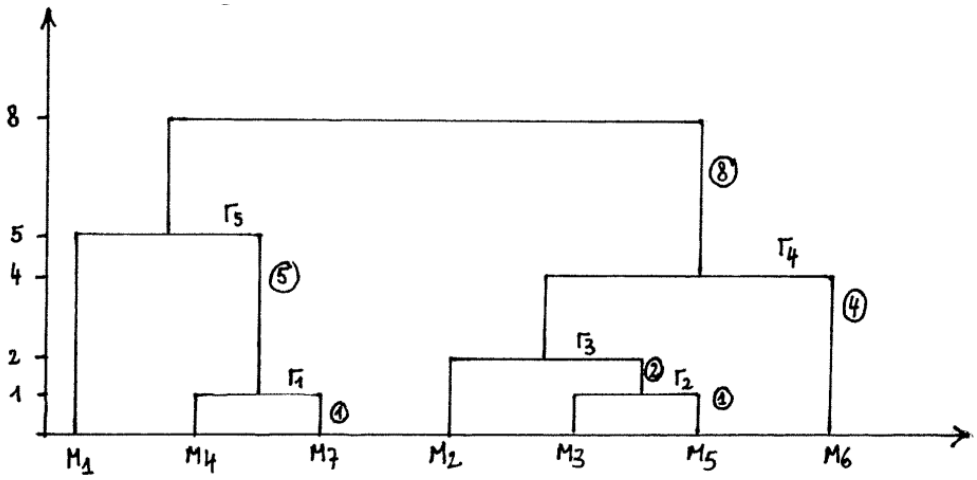


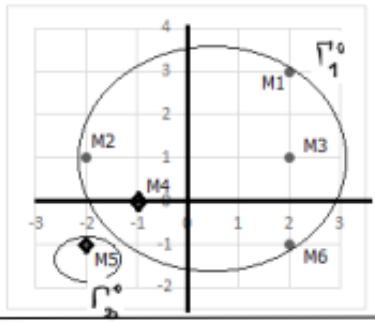
En regroupe M1 et Γ1 —— Γ5 = {M1,M4,M7}

	Γ4	Γ5
Γ4	0	8
Γ5		0

— *Le centre de gravité du nuage est*

$$G = \begin{pmatrix} \frac{2-2+2-1-2+2}{6} \\ \frac{3+1+1+0-1-1}{6} \end{pmatrix} = \begin{pmatrix} \frac{1}{6} \\ \frac{1}{2} \end{pmatrix}$$

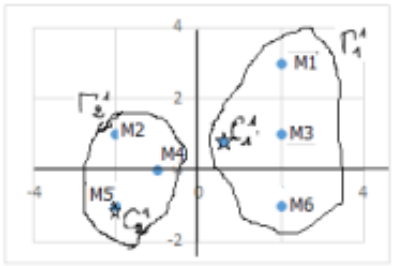




Première itération :

$C_1^0 = M_4$ $\quad C_2^0 = M_5$

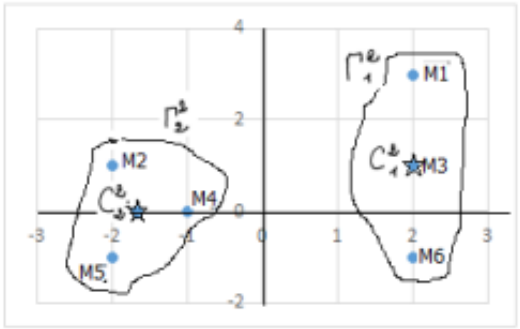
$\Gamma_1^0 = \{M_1, M_2, M_3, M_4, M_6\}$ $\quad \Gamma_2^0 = \{M_5\}$



Deuxième itération :

$C_1^1 = \begin{pmatrix} \frac{2-2+2-1+2}{5} \\ \frac{3+1+1+0-1}{5} \end{pmatrix} = \begin{pmatrix} 0,6 \\ 0,8 \end{pmatrix}$ $\quad C_2^1 = M5$

$\Gamma_1^1 = \{M_1, M_3, M_6\}$ $\quad \Gamma_2^1 = \{M_2, M_4, M_5\}$



Troisième itération :

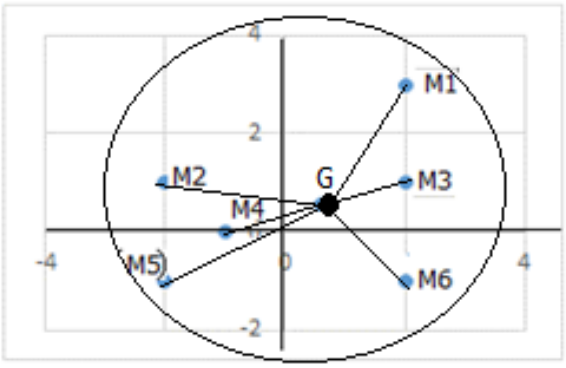
$C_1^2 = \begin{pmatrix} \frac{2+2+2}{3} \\ \frac{3+1-1}{3} \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} = M_3$ $\quad C_2^2 = \begin{pmatrix} \frac{2-1-2}{3} \\ \frac{1+0-1}{3} \end{pmatrix} = \begin{pmatrix} -1.667 \\ 0 \end{pmatrix}$

$\Gamma_1^2 = \{M_1, M_3, M_6\}$ $\quad \Gamma_2^2 = \{M_2, M_4, M_5\}$

La procédure est stoppé puisque la partition est resté inchangée entre la deuxième et la troisième itération .

— *L'inertie total*

$$I_{Total} = I_T = \frac{1}{6}\left[d^2(M_1,G)^2 + d^2(M_2,G)^2 + d^2(M_3,G)^2 + d^2(M_4,G)^2 + d^2(M_5,G)^2 + d^2(M\right.$$
$$= \frac{1}{6}\left[9.61 + 4.94 + 3.61 + 1.61 + 6.94 + 5.61\right] = \underline{\underline{5.39}};$$



— *Pour la première partition* $\Gamma_1^0 = (M_1, M_2, M_3, M_4, M_6$ *et* $\Gamma_2^0 = M_5, .$

$$\begin{aligned}
\mathbf{I_{intra}} &= \mathbf{I_{\Gamma_1^0}} + \mathbf{I_{\Gamma_2^0}} \\
&= \frac{1}{6}\left[d^2(M_1, C_1^1)^2 + d^2(M_2, C_1^1)^2 + d^2(M_3, C_1^1)^2 + d^2(M_4, C_1^1)^2 + d^2(M_6, C_1^1)^2\right] + 0 \\
&= \frac{1}{6}\left[6.8 + 6.8 + 2 + 3.2 + 5.2\right] = \underline{\underline{4}}
\end{aligned}$$

$$\mathbf{I_{inter}} = \frac{5}{6}d^2(C_1^1, G) + \frac{1}{2}\frac{5}{6}d^2(C_2^1 = M_5, G) = \underline{\underline{1.39}},$$

— *Pour la deuxième partition* $\Gamma_1^1 = (M_1, M_3, M_6)$ *et* $\Gamma_2^1 = (M_2, M_4, M_5), .$

$$\begin{aligned}
\mathbf{I_{intra}} &= \mathbf{I_{\Gamma_1^1}} + \mathbf{I_{\Gamma_2^1}} \\
&= \frac{1}{6}\left[d^2(M_1, C_1^2 = M_3)^2 + d^2(M_6, M_3)^2 + 0\right] + \frac{1}{6}\left[d^2(M_2, C_2^2)^2 + d^2(M_4, C_2^2)^2 + +d^2(M\right. \\
&= \frac{1}{6}(4+4) + \frac{1}{6}[\frac{10}{9} + \frac{4}{9} + \frac{10}{9}] = \underline{\underline{1.78}};
\end{aligned}$$

$$\begin{aligned}
\mathbf{I_{inter}} &= \frac{3}{6}d^2(C_1^2, G) + \frac{3}{6}d^2(C_2^2, G) \\
&= \frac{1}{2}(3.61) + \frac{1}{2}d^2(3.61) = \underline{\underline{3.612}};
\end{aligned}$$

On observe qu'a chaque itération l'inertie intra diminué (passe de 5.92 *à* 4 *puis* 1.78*) et l'inertie inter augmente (elle passe de* 0 *à* 1.39 *puis* 3.61 *) mais la somme des deux reste toujours égale l'inertie total* $(4 + 1.39 = 5.39, \qquad 1.78 + 3.61 = 5.39)$

3. *Classifier les points du nuage précédent par une classification hiérarchique ascendante (CAH) et représenter le dendrogramme.*

Bibliographie

[1] Bouroche, J.-M. et Saporta, G. (1987). *Analyse des Données. Que sais-je ?*, Paris.

[2] Durand, J.F. (2002) .*Calcul Matriciel et Analyse Factorielle des Données.Polycopier de cours*-Universitt'e Montpellier II -France.

[3] Escofier.B et J. pagès . (2008) .*Analyses factorielles simples et multiples,* Dunod, Paris, .

[4] Lebart L. Morineau A. Piron M. (1995). *Statistique exploratoire multidimensionnelle*, Dunod.Paris,

[5] Martin, A. (2003) . *L'analyse de données. Polycopier de cours ESNIETA*- ref 1463. .

[6] Martinez, Wendy L.(2011) *Exploratory data analysis with MATLAB* . – 2nd ed.Chapman & Hall

[7] Saporta,G. (2005). *Probabilités, analyses des données et statistiques* , Editions Technip.Paris,

[8] Sanders Lena,(1989). *L'anaLyse des données appliquée à la géographie.* Montpellier, G.I.P. RECLUS, .