Institutional Repository - Research Portal
Dépôt Institutionnel - Portail de la Recherche
researchportal.unamur.be
University of Namur

THESIS / THÈSE

MASTER EN SCIENCES MATHÉMATIQUES

Méthodes de détermination du nombre de classes pour des objets symboliques

Troclet, Jennifer

Award date:
2004

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

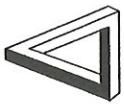
Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 02. May. 2026

FUNDP
Faculté des Sciences
Département de Mathématique

Rempart de la Vierge, 8
B-5000 Namur Belgique

Méthodes de détermination du nombre de classes pour des objets symboliques



Mémoire présenté pour l'obtention
du grade de
Licencié en Sciences Mathématiques
par

Jennifer TROCLET

Promoteur : Prof. A. Hardy

Année Académique 2003-2004

Je tiens à remercier tout particulièrement mon promoteur, Monsieur A. Hardy, qui m'a guidée et aidée dans l'élaboration de ce mémoire. Je voudrais aussi remercier Madame P. Lallemand qui m'a aidée à résoudre un grand nombre de problèmes informatiques.

Je remercie également l'ensemble des professeurs qui ont contribué à ma formation.

Enfin, je remercie ma famille et mes amis qui m'ont soutenue tout au long de mes études.

Résumé

Dans ce mémoire, nous nous intéressons à la détermination du nombre de classes pour des données symboliques décrites par des variables de type intervalle, multivaluées, modales et par des combinaisons de ces trois types. Nous adaptons les cinq meilleures méthodes de détermination du nombre de classes issues de l'étude de Milligan et Cooper au programme de classification symbolique Sclust ainsi qu'à quatre méthodes de classification hiérarchiques (saut minimum, lien complet, Ward et centroïde). Nous comparons les distances disponibles dans le module DISS du logiciel SODAS avec les distances plus classiques (L_1, L_2, de Hausdorff et de De Carvalho). Nous testons ces méthodes sur différents ensembles de données artificiels et réels et analysons les résultats obtenus.

Abstract

In this report, we are interested in the determination of the number of clusters for symbolic data described by interval, multi-valued and modal variables and by a combination of this three type. We adapt the five best methods of determination of the number of clusters stemmed from the study of Milligan and Cooper to the program of symbolic classification Sclust as to four hierarchical methods of classification (single linkage, complete linkage, Ward and centroid). We compare the distance available in DISS module of SODAS software and with the more classical distances (L_1, L_2, de Hausdorff et de De Carvalho). We test these methods on various artificial and real data sets and analyse the obtained results.

Table des matières

Introduction 5

I Théorie 6

1 Les données classiques 7
- 1.1 Introduction . . . 7
- 1.2 Types de variables [3] . . . 7
  - 1.2.1 Les variables quantitatives . . . 7
  - 1.2.2 Les variables qualitatives . . . 8
- 1.3 Vecteurs et matrices de données . . . 9
- 1.4 Exemple . . . 10

2 Les données symboliques 11
- 2.1 Introduction . . . 11
- 2.2 Les variables multivaluées et de type intervalle . . . 11
  - 2.2.1 Variable multivaluée [3] . . . 11
  - 2.2.2 Variable de type intervalle [3] . . . 12
  - 2.2.3 Variables multivaluées et intervalle par agrégation . . . 13
- 2.3 Variables modales [3] . . . 13
- 2.4 Résumé des types de données symboliques [3] . . . 14
- 2.5 Le tableau des données symboliques [3] . . . 15
- 2.6 Passage de données classiques à des données symboliques . . . 16
- 2.7 Dissimilarités entre objets . . . 18
  - 2.7.1 Le cas des données symboliques de type intervalle . . . 18
  - 2.7.2 Le cas des variables multivaluées . . . 19
  - 2.7.3 Le cas des variables modales . . . 20

3 La classification automatique 21
- 3.1 Introduction . . . 21
- 3.2 Structures classificatoires . . . 22
  - 3.2.1 Les partitions . . . 22
  - 3.2.2 Les hiérarchies . . . 22

3.3 Les méthodes de classification 20
3.3.1 Les méthodes hiérarchiques agglomératives 20
3.3.2 Les méthodes non-hiérarchiques ou de partitionnement 21
3.4 Les méthodes hiérarchiques agglomératives 21
3.4.1 La méthode du saut minimum [5] 21
3.4.2 La méthode du saut maximum [5] 21
3.4.3 La méthode du centroïde [5] 22
3.4.4 La méthode de Ward [6] 22
3.5 Classification de données symboliques 23

4 La méthode des nuées dynamiques **24**
4.1 Notion d'inertie 24
4.1.1 Inertie par rapport à un point [7] 24
4.1.2 Théorème de Huygens[7] 24
4.1.3 Inerties associées à une partition [7] 25
4.1.4 Décomposition des inerties sur les classes et les variables [7] 26
4.2 La méthode des nuées dynamiques dans le cas de données classiques [7] . . 27
4.2.1 Introduction 27
4.2.2 Principe général 27
4.2.3 Aspect formel 28
4.2.4 Le cas du centre de gravité 29
4.3 La méthode des nuées dynamiques dans le cas de données symboliques . . 31
4.3.1 Le cas des données de type intervalle 31
4.3.2 Le cas des variables multivaluées et modales 33

5 Méthodes de détermination du nombre de classes **35**
5.1 Introduction 35
5.2 Les méthodes de Milligan et Cooper 35
5.2.1 La méthode de Calinski et Harabasz (1974) [8] 35
5.2.2 La méthode de Duda et Hart [9] 37
5.2.3 La méthode du C-index [10] 38
5.2.4 La méthode Gamma [11] 38
5.2.5 La méthode de Beale [12] 39

6 Le module DISS **41**
6.1 Une nouvelle subdivision des objets symboliques 41
6.2 Mesures de dissimilarité entre objets booléens 42
6.2.1 La mesure de dissimilarité de Gowda et Diday (U-1) 42
6.2.2 L'approche d'Ichino et Yaguchi (U-2, U-3, U-4) 45
6.2.3 Mesures de dissimilarité de De Carvalho 48
6.2.4 Mesure de dissimilarité pour les BSO avec contraintes 52
6.3 Dissimilarités pour des distributions de probabilité 53
6.3.1 Mesures de divergence : Le cas général 54

6.3.2 Mesures de divergence : Cas spéciaux 55
6.3.3 Mesures de dissimilarité entre deux PSO 59
6.4 Le module DISS 61
6.4.1 Introduction 61
6.4.2 L'entrée du module DISS 61

7 Le programme SCLUST **72**
7.1 Introduction 72
7.2 Brève description de SCLUST 73
7.2.1 Le fichier sodas 73
7.2.2 Le fichier sclust.h 76
7.2.3 Le fichier calcul_scluster.cpp 76
7.2.4 La fonction main_scluster 78
7.2.5 Le fichier listing 79
7.3 Adaptation du programme SCLUST 81
7.3.1 Construction des hiérarchies de partitions 82
7.3.2 Méthodes de détermination du nombre de classes de Milligan et Cooper 82

II Applications **83**

8 Introduction **84**

9 Présentation des résultats **86**

10 Les variables de type intervalle **88**
10.1 Introduction 88
10.2 Analyse des distances présentes dans DISS 88
10.2.1 Données avec trois classes hypersphériques 89
10.2.2 Données avec deux classes allongées 98
10.2.3 Conclusion 105
10.3 Nouveau programme 106
10.3.1 Résultats obtenus pour les données avec trois classes hypersphériques 107
10.3.2 Résultats obtenus pour les données avec deux classes allongées . . . 108
10.4 Comparaison des résultats sur diverses jeux de données 110
10.4.1 Données avec trois classes hypersphériques 110
10.4.2 Données avec deux classes allongées 112
10.4.3 Données sans structure 114
10.4.4 Données avec deux classes emboîtées 119
10.4.5 Données basées sur les formes de Breiman [1] 124
10.4.6 Températures de villes chinoises 132
10.5 Conclusion 138

11 Les variables multivaluées **139**
11.1 Introduction 139
11.2 Comparaison des distances implémentées dans DISS pour le jeu de données "ANIMAUX" 139
11.2.1 Informations sur le jeu de données 139
11.2.2 Comparaison 140
11.2.3 La distance U-1 141
11.2.4 La distance U-2 142
11.2.5 Les distances U-3, U-4, SO-1 et S0-2 143
11.2.6 La distance SO-3 143
11.2.7 La distance SO-4 143
11.2.8 Conclusion 144
11.3 Comparaison de notre méthode avec celle de Séverine Collès 145
11.3.1 "Données Animaux" 145
11.3.2 Boucles mérovingiennes datant du 6-8ème siècle après Jésus-Christ [2] 151

12 Les variables modales **162**
12.1 Introduction 162
12.2 Comparaison de nos distances avec celles utilisées par Séverine Collès . . . 163
12.2.1 Magasins e-Fashion 163
12.2.2 Conclusion 175
12.2.3 Consommation [2] 177
12.2.4 Conclusion 199

13 Combinaisons de variables **200**
13.1 Introduction 200
13.2 Voitures 200
13.2.1 Informations sur le jeu de données 200
13.2.2 Résultats obtenus avec le module DISS 202
13.2.3 Analyse 205

Conclusion **207**

Introduction

De nos jours, il est devenu essentiel de pouvoir traiter de grands ensembles de données. En effet, les progrès réalisés dans le domaine informatique permettent la gestion et le traitement de données à grande échelle. Les informations récoltées peuvent parfois être imprécises ou plus "complexes" que les données classiques que l'on traite habituellement ce qui a conduit à l'essor des données dites symboliques. C'est pour ces raisons que depuis plusieurs années il y a nécessité d'étendre les méthodes d'analyse des données classiques à ces types de données récentes.

Dans ce mémoire, nous nous intéresserons tout particulièrement à la classification des données symboliques. Ces nouvelles méthodes ont des applications dans de nombreux domaines comme la finance, le traitement d'images, l'industrie,... La classification est utilisée pour la résolution du problème de la *division d'une population donnée d'individus ou d'objets ou d'individus semblables. Plusieurs méthodes de classification existent et permettent d'obtenir une partition des individus, décrits par un ensemble de variables, en un certain nombre de classes généralement fixé a priori.* Parfois, ce nombre correspond au nombre de classes composant la partition la plus naturelle.

A coté de ce problème de classification, nous étudierons des méthodes de détermination du nombre de classes. Celles-ci sont très intéressantes car lors de l'étude de données réelles, le nombre de classes formées par les individus est inconnu. Ces méthodes ont été étudiées pour des données classiques et ont été modifiées il y a peu pour des données symboliques. Jusqu'ici ces méthodes trouvaient réponse pour des données décrites par un seul type de variable symbolique. Le but de ce chapitre est d'adapter ces méthodes pour des combinaisons de variables symboliques.

Les deux premiers chapitres présentent les données classiques et les données symboliques. Nous exposerons ensuite des méthodes de classification classiques et plus particulièrement la méthode des nuées dynamiques. Nous présenterons par la suite les cinq premières méthodes de détermination du nombre de classes issues du classement de Milligan et Cooper réalisé en 1988. Puis, nous étudierons le module que nous avons utilisé afin d'obtenir la matrice de distance utilisée dans notre programme. Nous détaillerons alors le programme de classification symbolique Sclust. Enfin, nous appliquerons les méthodes de détermination du nombre de classes adaptées aux objets décrits par n'importe quel type ou combinaison de variables et comparerons si possible les résultats avec des méthodes utilisées antérieurement.

Première partie

Théorie

Chapitre 1

Les données classiques

1.1 Introduction

Considérons :
- $E = \{x_1, \ldots, x_n\}$, un ensemble des n individus ;
- $Y_1, \ldots, Y_p$, p variables qui caractérisent chaque individu ;
- $\mathcal{Y}_j$, l'ensemble des valeurs prises par la variable Y_j $(j \in \{1, \ldots, p\})$ appelé espace ou domaine d'observation de la variable Y_j.

Une **variable classique** est définie par

$$Y_j : E \to \mathcal{Y}_j$$

$$x_k \rightsquigarrow Y_j(x_k) = x_{kj}$$

où x_{kj} est la valeur observée de la variable j pour l'individu x_k.

Toutes ces valeurs sont placées dans une matrice de données

$$\widetilde{X} = (x_{kj})_{n \times p}$$

1.2 Types de variables [3]

Nous distinguons deux types de variables :

1. les variables quantitatives et
2. les variables qualitatives.

Nous considérerons différentes mesures de proximité δ_j entre deux individus $x, y \in E$ tel que $\delta_j(x, y)$ est la mesure de "dissimilarité" entre ces deux éléments.
Remarque : Par la suite, l'indice de la variable sera omis.

1.2.1 Les variables quantitatives

Une variable Y est dite **quantitative** si l'espace d'observation $\mathcal{Y}$ est tel que $\mathcal{Y} \subseteq \mathbb{R}$.

Variable quantitative continue

Une variable quantitative Y est **continue** si elle prend un nombre infini non dénombrable de valeurs dans $\mathbb{R}$.
Dans ce cas, nous pouvons avoir :
- $\mathcal{Y} = \mathbb{R}$,
- $\mathcal{Y} = \mathbb{R}^+$, ou encore
- $\mathcal{Y} = [a\ ,\ b] = \{x \in \mathbb{R} | a \leq x \leq b\}$ où $-\infty < a < b < \infty$.

La mesure de proximité $\delta(x, y)$ entre deux éléments $x, y \in \mathcal{Y}$ est fournie par la distance euclidienne, i.e.

$$\delta(x, y) = |x - y|$$

Variable quantitative discrête

Une variable quantitative Y est **discrête** si l'espace d'observation $\mathcal{Y}$ contient un nombre fini ou un nombre infini dénombrable de valeurs $\xi_i \in \mathbb{R}$.
Formellement, nous avons $\mathcal{Y} = \{\xi_1, \ldots, \xi_N\} \subset \mathbb{R}$ ou $\mathcal{Y} = \{\xi_1, \xi_2, \ldots\} \subset \mathbb{R}$.

1.2.2 Les variables qualitatives

Une variable Y est dite **qualitative** ou **catégorique** si le nombre de valeurs de l'espace d'observation $\mathcal{Y}$ est fini et si les éléments de $\mathcal{Y}$, appelés catégories, ne portent aucune structure.

Variable nominale

Une variable qualitative est **nominale** si elle a des modalités distinctes les unes des autres, mais sans structure interne, c'est-à-dire, sans possibilité d'ordre ou de calcul entre elles.
Dans ce cas, pour deux catégories $x, y \in \mathcal{Y}$, nous ne pouvons seulement distinguer que $x = y$ ou $x \neq y$.
La mesure de proximité $\delta(x, y)$ entre deux catégories $x, y \in \mathcal{Y}$ est définie par :

$$\delta(x, y) = \begin{cases} 1 & si \quad x = y; \\ 0 & sinon \end{cases}$$

Dans le cas particulier des variables **binaires** ou **dichotomiques**, l'espace d'observation $\mathcal{Y}$ ne comprend que deux modalités, codées habituellement par 0 et 1.
Lorsque la variable nominale considérée a plus de deux modalités, les s catégories peuvent être codées $0, 1, 2, \ldots, s-1$. Cependant, aucune opération arithmétique ne peut être définie avec ces codes.

Variable ordinale

Une variable qualitative est **ordinale** si elle a des modalités qui peuvent être hiérarchisées entre elles, aucun calcul ne peut cependant être défini.
Cela revient à dire que l'espace d'observation $\mathcal{Y}$ est muni d'un ordre total $\prec$ défini tel que $\forall x, y \in \mathcal{Y},\ x \neq y$, nous avons soit $x \prec y$ soit $y \prec x$.

Tout comme pour les variables nominales, les différentes catégories peuvent être codées de sorte que $\mathcal{Y} = \{0, 1, 2, \ldots, s-1\}$ où s est le nombre de modalités prises par cette variable. Il est bien entendu proscrit de définir des opérations arithmétiques avec ces codes. Cependant, ils peuvent servir à la définition d'une mesure de proximité par :

$$\delta(x, y) = |x - y| \ \forall x, y \in \mathcal{Y}$$

qui représente en fait le nombre de catégories de $\mathcal{Y}$ strictement comprises entre x et y selon l'ordre total imposé par $\prec$.

1.3 Vecteurs et matrices de données

Comme précédemment, considérons l'ensemble des n individus $E = \{x_1, x_2, \ldots, x_n\}$ caractérisés par p variables notées $Y_1, \ldots, Y_p$; $\mathcal{Y}_j$ représente l'espace d'observation de la variable Y_j où $j = 1, \ldots, p$.

Nous désignons par X le vecteur des p variables $Y_1, \ldots, Y_p$:

$$X = \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix} \in \mathcal{X} = \bigotimes_{j=1}^{p} \mathcal{Y}_j$$

où $\bigotimes_{j=1}^{p} \mathcal{Y}_j$ est le produit cartésien des p espaces d'observations $\{\mathcal{Y}_1, \ldots, \mathcal{Y}_p\}$.

La valeur ou la catégorie de Y_j observée pour l'individu $x_k \in E$ est notée $x_{kj} = Y_j(x_k)$.

Pour chaque individu $x_k \in E$, nous pouvons représenter les p observations $\{x_{k1}, \ldots, x_{kp}\}$ dans un vecteur-colonne p-dimensionnel :

$$x_k = X(x_k) = \begin{pmatrix} x_{k1} \\ x_{k2} \\ \vdots \\ x_{kp} \end{pmatrix} \in \mathcal{X} = \bigotimes_{j=1}^{p} \mathcal{Y}_j$$

En prenant en compte les n individus, nous obtenons la matrice des données classiques :

$$\widetilde{X} = (x_{kj})_{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_p \end{pmatrix} = (y_1, \dots, y_p)$$

tel que la k-ième ligne x'_k contient les données observées pour l'individu k et la j-ième colonne y_j représente les valeurs prises par la variable Y_j sur l'ensemble des individus.

1.4 Exemple

Soit $E = 1, 2, 3, 4, 5$ un ensemble de cinq pays pour lesquels quatre variables ont été mesurées :

- Y_1 = la capitale (qualitative nominale).
- Y_2 indique si le pays utilise l'euro ou non. Cette variable vaut 1 si oui et 0 sinon (qualitative nominale binaire).
- Y_3 = le nombre d'individus peuplant le pays (quantitative discrète)
- Y_4 indique le type linguistique du pays (unilingue, bilingue, trilingue) (qualitative ordinale).

La matrice des données $\widetilde{X}$ correspondante est de la forme suivante :

	Y_1	Y_2	Y_3	Y_4
1	Bruxelles	1	10 140 000	trilingue
2	Paris	1	58 680 000	unilingue
3	Luxembourg	1	422 000	trilingue
4	Copenhague	0	5 270 000	unilingue
5	Royaume-Uni	0	58 650 000	unilingue

Remarque : Cette matrice de données est un **tableau mixte** car elle contient des variables qui ne sont pas toutes de la même nature.

Chapitre 2

Les données symboliques

2.1 Introduction

Nous allons introduire trois types de variable :
- les variables multivaluées ;
- les variables de type intervalle ;
- les variables modales.

L'ensemble des objets E peut-être défini de deux façons différentes :

1. un ensemble $E = \{x_1, \ldots, x_n\}$ d'individus appelés **objets du premier ordre** ;
2. un ensemble $E = \{C_1, C_2, \ldots\}$ de classes $C_i \subseteq E$ d'individus appelées **objets du second ordre**.

Des objets d'ordre supérieur peuvent être définis de manière similaire par des étapes d'agrégation successives.

2.2 Les variables multivaluées et de type intervalle

2.2.1 Variable multivaluée [3]

La variable Y, dont l'espace d'observation est $\mathcal{Y}$, est dite à valeurs dans un ensemble $\mathcal{B}$ lorsque :

$$Y : E \to \mathcal{B}$$

$$x_k \rightsquigarrow Y(x_k)$$

où $\mathcal{B} = \mathcal{P}(\mathcal{Y}) = \{U \neq \varnothing | U \subseteq \mathcal{Y}\}$ est l'*ensemble des parties de l'espace d'observation* $\mathcal{Y}$.
Une variable Y est dite **multivaluée** lorsque les valeurs $Y(x_k)$ sont toutes des sous-ensembles finis de $\mathcal{Y}$, c'est-à-dire

$$|Y(x_k)| < \infty, \ \forall x_k \in E.$$

Une variable Y est dite **multivaluée catégorique** si $\mathcal{Y}$ a un nombre fini de catégories et donc

$$|Y(x_k)| < \infty, \ \forall x_k \in E.$$

Une variable Y est dite **multivaluée quantitative** si les valeurs $Y(x_k)$ sont des ensembles finis de nombres réels, c'est-à-dire :

$$Y(x_k) \subset \mathbb{R} \ et \ |Y(x_k)| < \infty, \forall x_k \in E.$$

Exemple 2.2.1 *Variable multivaluée catégorique et quantitative*
Considérons :

- $E = \{Opel, Peugeot, VW, Mercedes\}$
- $Y_1 =$ *les modèles proposés par différentes marques de voitures*
  $\mathcal{Y}_1 =$ { *corsa, classe A, 106, omega, SLK, vectra, 206, Beetle, 307, polo, golf, Classe C* }
- $Y_2 =$ *le prix de base en euro des différents modèles*
  $\mathcal{Y}_2 = \mathbb{R}^+$

Voici les résultats pour les cinq marques de voitures de l'ensemble E :

	Y_1	Y_2
Opel	*{Corsa, Omega, Vectra}*	*{ 10900, 21480, 26300 }*
Peugeot	*{106, 206, 307}*	*{8950, 9900, 14460}*
VW	*{Polo, Golf, Beetle}*	*{ 10960, 15470, 17170}*
Mercedes	*{ Classe A, Classe C, SLK }*	*{ 17061, 26015, 34001 }*

Dans cet exemple, Y_1 est une variable multivaluée catégorique et Y_2 est une variable multivaluée quantitative.

◇

2.2.2 Variable de type intervalle [3]

Une variable est de type **intervalle** si $\forall x_k \in E$, l'ensemble $Y(x_k)$ est un intervalle borné et fermé de $\mathbb{R}$.

Dans ce cas, $\mathcal{B}$ est l'ensemble des intervalles fermés bornés de $\mathbb{R}$, on a donc que $\mathcal{Y} = \mathcal{I}$.

Exemple 2.2.2 *Variable de type intervalle*
Soient :

- $E = \{$*élèves d'une école*$\}$
- $Y =$ *le temps de sommeil moyen de chaque élève (en heures)*
- $\mathcal{Y} = \{[a,b] | \ a, b \in \mathbb{R}^+, 0 \leq a, b \leq \infty\}$

Nous pouvons obtenir les résultats suivants : $Y(x_k) = [7,8], Y(x_\ell) = [8,9], Y(x_i) = [9,10]$ *où* $x_k, x_\ell, x_i \in E$

◇

2.2.3 Variables multivaluées et intervalle par agrégation

Supposons que
- $\Omega = \{x_1, \ldots, x_n\}$ est l'ensemble des objets du premier ordre ;
- $\widetilde{Y}$ est une variable univaluée classique et
- $E = \{C_1, \ldots, C_m\}$ est l'ensemble des classes $C_i \subseteq \Omega$, appelés objets du second ordre.

Nous cherchons à caractériser le comportement de ces classes par rapport à la variable $\widetilde{Y}$.

Une solution est de définir une variable "globale" ou "agrégée" Y qui spécifie les valeurs prises par $\widetilde{Y}$ sur les classes C_i.

Exemple 2.2.3 *Variable multivaluée obtenue par agrégation*

- $\Omega = \{$*élèves d'une école*$\}$
- $E = \{C_1, \ldots, C_m\} = \{m$ *classes*$\}$
- $\widetilde{Y}(x_k) =$ *le temps de sommeil moyen de l'individu* $x_k \in E$ *(en heures).*

La description de la classe C_i *est*

$$Y(C_i) = \{7.45, 8.30, 8.00, 9.15, 7.55, 8.10, 9.30, 8.25, 7.50, 9.00, 8.55\}$$

c'est-à-dire l'ensemble des temps de sommeil moyen des élèves de la classe C_i.

◇

Exemple 2.2.4 *Variable intervalle obtenue par agrégation*

Reprenons les données de l'exemple précédent.

Nous pouvons décrire la classe C_i *par* $Y(C_i) = [\alpha, \beta]$ *où*

$$\alpha = \min_{\omega \in C_i} \{\widetilde{Y}(\omega)\}$$

$$\beta = \max_{\omega \in C_i} \{\widetilde{Y}(\omega)\}$$

et par conséquent, $Y(C_i) = [7.45, 9.30]$

◇

2.3 Variables modales [3]

Une variable **modale** sur $E = \{x_1, \ldots, x_n\}$ dont l'espace d'observation est $\mathcal{Y}$ est une fonction

$$Y(x_k) = (U(x_k), \pi_k), \forall x_k \in E$$

où

- π_k est, par exemple, une distribution de fréquence sur les observations Y,
- $U(x_k) \subseteq \mathcal{Y}$ est le support de π_k dans le domaine $\mathcal{Y}$.

Exemple 2.3.1 *Variable modale*

Soient
- $\Omega = \{x_1, \ldots, x_{100}\}$ *des individus belges qui louent un appartement,*
- $\widetilde{Y}$ *une variable univaluée classique qui indique le loyer payé par ces personnes et*
- $C = \{1, \ldots, 10\}$ *la classe des dix premiers individus.*

$\widetilde{Y}(C)$ *prend les valeurs suivantes : 356, 421, 562, 404, 625, 460, 387, 550, 475, 575.*

La variable modale Y qui décrit le loyer payé dans la classe C peut avoir la forme d'un histogramme

$$\widetilde{Y}(C) = \{(]300, 400], \frac{2}{10}), (]400, 500], \frac{3}{10}), (]500, 600]\frac{3}{10}), (]600, 700], \frac{2}{10})\}$$

◇

2.4 Résumé des types de données symboliques [3]

Une **variable symbolique** d'espace d'observation $\mathcal{Y}$ est définie comme suit :

$$Y : E \to \mathcal{B} \quad \forall x_k \in E$$

$$x_k \rightsquigarrow Y(x_k)$$

où $\mathcal{B} = \mathcal{P}(\mathcal{Y}) = \{U \neq \varnothing | U \subseteq \mathcal{Y}\}$.

1. Si $\mathcal{B} = \mathcal{Y}$, nous sommes dans le cas d'une variable classique univaluée.
2. Y est à valeurs dans un ensemble $\mathcal{B}$
   si $Y(x_k) \subseteq \mathcal{Y}$, $\forall x_k \in E$, ce qui revient à considérer $\mathcal{B} = \mathcal{P}(\mathcal{Y})$.
3. Y est une variable de type intervalle
   si, $\forall x_k \in E, Y(x_k) = [\alpha, \beta]$ est un intervalle de $\mathcal{Y}$ et donc $\mathcal{B}$ est l'ensemble Jdes intervalles fermés bornés de $\mathcal{Y}$.
4. Y est une variable multivaluée (catégorique ou quantitative)
   si $Y(x_k) \subseteq \mathcal{Y}$ et $|Y(x_k)| < \infty, \forall x_k \in E$.
5. Y est une variable modale d'espace d'observation $\mathcal{Y}$
   si, pour chaque $x_k \in E, Y(x_k) = \pi_a$ est une mesure non négative sur $\mathcal{Y}$, habituellement une distribution de fréquence, de probabilité ou un poids, d'où $\mathcal{B} = \mathcal{M}(Y)$.

Ceci nous permet de considérer les données symboliques comme une extension des données classiques.

2.5 Le tableau des données symboliques [3]

Considérons un ensemble de base $\Omega = \{x_1, \ldots, x_n\}$ de n individus. Considérons un ensemble d'objets $E = \{x_1, \ldots, x_N\}$. E peut être :

- l'ensemble des n individus, i.e. $E = \Omega = \{x_1, \ldots, x_n\}(N = n)$;
- un sous-ensemble $E \subset \Omega$, i.e. un échantillon de Ω $(N < n)$;
- un sous-ensemble $E = \{C_1, \ldots, C_m\}$ de classes $C_1, \ldots, C_m \subseteq \Omega$ d'individus $x_k \in \Omega(N = n)$. Dans ce cas, E est un ensemble d'objets du second ordre.

Considérons aussi p variables symboliques $Y_1, \ldots, Y_p$, où Y_j a pour espace d'observation $\mathcal{Y}_j$, ainsi qu'un objet $x_k \in E$.

Notons par $X(x_k) \equiv (Y_1(x_k), \ldots, Y_p(x_k))'$ le vecteur des variables symboliques déterminée pour $x_k \in E$.

Ainsi, chaque objet $x_k \in E$ peut être décrit par un vecteur de données symboliques :

$$\widetilde{x}_k = X(x_k) = \begin{pmatrix} x_{k1} \\ \vdots \\ x_{kp} \end{pmatrix} \in \mathcal{X} = \bigotimes_{j=1}^{p} \mathcal{B}_j$$

où

- $x_{kj} = Y_j(x_k) \in \mathcal{B}_j$ est la valeur de la j^{me} variable symbolique Y_j pour l'individu $x_k (j = 1, \ldots, p)$, et
- $\bigotimes_{j=1}^{p} \mathcal{B}_j$ est le produit cartésien des p espaces d'observation $\mathcal{B}_1, \ldots, \mathcal{B}_p$.

Toutes les données peuvent être compilées dans une **matrice de données symboliques**

$$\underline{X} = \begin{pmatrix} \widetilde{x}_1{}' \\ \vdots \\ \widetilde{x}_N' \end{pmatrix} = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{N1} & \ldots & x_{Np} \end{pmatrix} = (x_{ij})_{N \times p}$$

où la cellule x_{ij} peut contenir soit un ensemble, soit un intervalle ou encore un histogramme.

La k^{me} ligne de ce tableau est la description symbolique de l'élément $x_k \in E$.

Exemple 2.5.1 *Tableau de données symboliques*

Soient :
- $E = \{p_1, p_2, p_3, p_4\}$ *un ensemble de quatres provinces belges wallones* $(N = 4)$.
- Y_1 = *la température moyenne (minimale et maximale de ces dix dernières années en degré celsius).*
  $\mathcal{B}_1 = \mathcal{I} = \{[\alpha, \beta] = [min, max]$ *tel que* $-\infty < \alpha < \beta < \infty\}$.
  Y_1 *est donc une variable de type intervalle.*
- Y_2 = *les parcs d'attractions et de détente de la province.*
  $\mathcal{Y}_2 = \{$ *Walibi, Telecoo, Monde Sauvage, Pardisio, Barvaux...*$\}$.
  $\mathcal{B}_3 = \mathcal{P}(\mathcal{Y}_3)$. Y_2 *est une variable multivaluée catégorique.*
- Y_3 = *le pourcentage de voix obtenu par le PS, le MR, le CDH et l'ensemble des autres partis aux élections législatives 2003 dans la province.*
  $\mathcal{Y}_3 = \{PS, MR, CDH, autres\}$.
  $\mathcal{B}_3$ *est l'ensemble des distributions de fréquences sur* $\mathcal{Y}_3$.

Nous obtenons la matrice suivante :

p_1	$[-3.8, 28.0]$	$\{Paradisio\}$	$\{(MR, \frac{24}{100}), (CDH, \frac{13}{100}), (PS, \frac{41}{100}), (autres, \frac{22}{100})\}$
p_2	$[-5.1, 27.1]$	$\{Barvaux,$ *Grotte de Han* $\}$	$\{(MR, \frac{31}{100}), (CDH, \frac{18}{100}), (PS, \frac{32}{100}), (autres, \frac{19}{100})\}$
p_3	$[-3.8, 28]$	$\{$ *Telecoo, Monde sauvage* $\}$	$\{(MR, \frac{32}{100}), (CDH, \frac{24}{100}), (PS, \frac{30}{100}), (autres, \frac{14}{100})\}$
p_4	$[-3.9, 27.2]$	$\{$ *Walibi* $\}$	$\{(MR, \frac{44}{100}), (CDH, \frac{15}{100}), (PS, \frac{23}{100}), (autres, \frac{18}{100})\}$

$\diamond$

2.6 Passage de données classiques à des données symboliques

Soit E un ensemble d'objets décrits par p variables classiques $\widetilde{Y}_1, \ldots, \widetilde{Y}_p$ d'espaces d'observation $\mathcal{Y}_1, \ldots, \mathcal{Y}_p$.

Décrivons les classes C_i par des variables symboliques $Y_1, \ldots, Y_p$ de telle façon que $Y_j(C_i)$ caractérise l'ensemble $\{\widetilde{Y}_j(x_k) | x_k \in C_i \subseteq \mathcal{Y}_j\}$ des valeurs observées pour $\widetilde{Y}_j$ à l'intérieur de la classe C_i.

Exemple 2.6.1 *Exemple de passage de données classiques à des variables symboliques*

Prenons une matrice de données $\widetilde{X} = (x_{kj})$ *de 14 voitures et trois variables univaluées classiques :*
- $\widetilde{Y}_1$ = *la marque de la voiture (Opel, Mercedes, VW, Peugeot) ;*
- $\widetilde{Y}_2$ = *le modèle de l'auto(Astra, Classe A, 106, Polo, Golf, ...) ;*
- $\widetilde{Y}_3$ = *le prix de base de la voiture.*

code voiture k	*marque* $\widetilde{Y}_1$	*modèle* $\widetilde{Y}_2$	*prix* $\widetilde{Y}_3$
1	*VW*	*Polo*	*10960*
2	*Peugeot*	*307*	*14460*
3	*Opel*	*Corsa*	*10900*
4	*VW*	*Beetle*	*17170*
5	*Opel*	*Vectra*	*21480*
6	*Mercedes*	*Classe A*	*17061*
7	*VW*	*Golf 5*	*15470*
8	*Opel*	*Omega*	*26300*
9	*Peugeot*	*106*	*8950*
10	*VW*	*Bora*	*16150*
11	*Peugeot*	*206*	*9990*
12	*Opel*	*Astra*	*15400*
13	*Mercedes*	*Classe C*	*26015*
14	*Mercedes*	*SLK*	*34001*

Nous allons considérer les groupes de voitures de marque VW, Peugeot, Opel et Mercedes. Ces groupes sont repectivement définis comme suit :

$$C_1 = \{1, 4, 7, 10\}$$

$$C_2 = \{2, 9, 11\}$$

$$C_3 = \{3, 5, 8, 12\}$$

$$C_4 = \{6, 13, 14\}$$

Nous montrons ci-après la matrice de données symboliques obtenues après avoir agrégé les individus en quatre classes.

code voiture k	*marque* $\widetilde{Y}_1$	*modèle* $\widetilde{Y}_2$	*prix* $\widetilde{Y}_3$
C_1	*VW*	*{ Polo, Beetle, Golf 5, Bora}*	*[10960,17170]*
C_2	*Peugeot*	*{ 307, 106, 206}*	*[8950,14460]*
C_3	*Opel*	*{Corsa, Vectra, Omega, Astra}*	*[10900,26300]*
C_4	*Mercedes*	*{Classe A, Classe C, SLK}*	*[17061,34001]*

$\diamond$

2.7 Dissimilarités entre objets

2.7.1 Le cas des données symboliques de type intervalle

Considérons une matrice de données symboliques $\underline{X}$ composée de n objets décrits par p variables de type intervalle. Nous avons :

$$\underline{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

où $x_{kj} = Y_j(x_k) = [\alpha_{kj}, \beta_{kj}]$ est la description de la j$^{\text{ème}}$ composante de l'objet $x_k \in E$.

Nous définissons une mesure de dissimilarité sur E à partir de p indices de dissimilarités sur les $\mathcal{B}_j$.

$$\delta_j : \mathcal{B}_j \times \mathcal{B}_j \to \mathbb{R}^+$$
$$(x_{kj}, x_{lj}) \rightsquigarrow \delta_j(x_{kj}, x_{lj})$$

A partir de deux intervalles $x_{kj} = [\alpha_{kj}, \beta_{kj}]$ et $x_{lj} = [\alpha_{lj}, \beta_{lj}]$, nous pouvons définir trois types de distances :

1. La distance de Haussdorff :

$$\delta_j(x_{kj}, x_{lj}) = max\{|\alpha_{kj} - \alpha_{lj}|, |\beta_{kj} - \beta_{lj}|\}$$

   Cette distance prend le maximum entre la valeur absolue de la différence des bornes inférieures des deux intervalles et la valeur absolue de la différence de ses bornes supérieures.

2. La distance L_1 :

$$\delta_j(x_{kj}, x_{lj}) = |\alpha_{kj} - \alpha_{lj}| + |\beta_{kj} - \beta_{lj}|$$

   Il s'agit de la somme des valeurs absolues des différences entre les bornes inférieures et les bornes supérieures.

3. La distance L_2 :

$$\delta_j(x_{kj}, x_{lj}) = (\alpha_{kj} - \alpha_{lj})^2 + (\beta_{kj} - \beta_{lj})^2$$

   qui est la somme des carrés des différences entre les bornes inférieures et supérieures.

Afin de se ramener à une dissimilarité sur l'ensemble E des objets, on définit :

$$d : E \times E \to \mathbb{R}^+$$
$$(x_k, x_l) \rightsquigarrow d(x_k, x_l) = \left(\sum_{j=1}^{p} \delta_j^2(x_{kj}, x_{lj}) \right)^2$$

où δ_j est la dissimilarité définie précédemment.

2.7.2 Le cas des variables multivaluées

Considérons $E = \{x_1, \ldots, x_n\}$ un ensemble de n objets décrits par p variables multivaluées $Y_1, \ldots, Y_p$ dont les espaces d'observations sont respectivement $\mathcal{Y}_1, \ldots, \mathcal{Y}_p$. $Y_j(x_k)$ est donc un ensemble de catégories de $\mathcal{Y}_j$.

Soit m_j le nombre de catégories prises par Y_j, c'est-à-dire $m_j = |\mathcal{Y}_j|$.

La fréquence $q_{j,x_k}(c_s)$ associée à chaque catégorie $c_s (s = 1, 2, \ldots, m_j)$ de Y_j pour l'objet x_k est donnée par

$$q_{j,x_k}(c_s) = \begin{cases} \frac{1}{|Y_j(x_k)|} & si \ \ c_s \in Y_j(x_k) \\ 0 & sinon. \end{cases}$$

Ainsi, la représentation symbolique de chaque objet $x_k \in E$ est un vecteur de dimension $m_1 + \ldots + m_p$, c'est-à-dire

$$x_k = ((q_{1,x_k}(c_1), \ldots, q_{1,x_k}(c_{m_1})), \ldots, (q_{p,x_k}(c_1), \ldots, q_{p,x_k}(c_{m_p})).$$

La matrice originale, $\underline{X} = (Y_j(x_k))$, est transformée en une matrice de fréquence $\widetilde{X}$ de dimension $n \times (m_1 + \ldots + m_p)$:

	Y_1			...	Y_p		
	1	...	m_1	...	1	...	m_p
x_1	$q_{1,x_1}(c_1)$	...	$q_{1,x_1}(c_{m_1})$	...	$q_{p,x_1}(c_1)$	...	$q_{p,x_1}(c_{m_p})$
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\vdots$
x_k	$q_{1,x_k}(c_1)$	...	$q_{1,x_k}(c_{m_1})$	...	$q_{p,x_k}(c_1)$	...	$q_{p,x_k}(c_{m_p})$
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\vdots$
x_n	$q_{1,x_n}(c_1)$	...	$q_{1,x_n}(c_{m_1})$	...	$q_{p,x_n}(c_1)$	...	$q_{p,x_n}(c_{m_p})$

où $\forall x_k \in E$, et $\forall j \in \{1, \ldots, p\}$, $\sum_{i=1}^{m_j} q_{j,x_k}(c_i) = 1$.

Grâce à cette matrice de fréquences, nous pouvons définir une mesure de dissimilarité sur E à partir de p indices de dissimilarités sur les $\mathcal{B}_j$. Considérons

$$\delta_j : \mathcal{B}_j \times \mathcal{B}_j \to \mathbb{R}^+$$
$$(x_{kj}, x_{lj}) \rightsquigarrow \delta_j(x_{kj}, x_{lj}).$$

Nous pouvons alors définir trois distances :

1. La distance L_1 :

$$\delta_j(x_{kj}, x_{lj}) = \sum_{i=1}^{|\mathcal{Y}_j|} |x_{kj}^{(i)} - x_{lj}^{(i)}|.$$

où x_{kj} est la fréquence prise par la variable j pour l'individu x_k concernant la modalité i, c'est-à-dire $x_{kj}^{(i)} = q_{j,x_k}(c_i)$

2. La distance L_2 :

$$\delta_j(x_{kj}, x_{lj}) = \sum_{i=1}^{|\mathcal{Y}_j|}(x_{kj}^{(i)} - x_{lj}^{(i)})^2.$$

3. La distance de De Carvalho :

$$\delta_j(x_{kj}, x_{lj}) = \sum_{i=1}^{|\mathcal{Y}_j|}(\gamma x_{kj}^{(i)} + \gamma' x_{lj}^{(i)})^2.$$

où

$\diamond$ $\gamma = \begin{cases} 1 & si \quad c_i \in Y_j(x_k) \text{ et } c_i \notin Y_j(x_l) \\ 0 & sinon \end{cases}$

$\diamond$ $\gamma' = \begin{cases} 1 & si \quad c_i \notin Y_j(x_k) \text{ et } c_i \in Y_j(x_l) \\ 0 & sinon \end{cases}$

On combine les p indices de dissimilarité $\delta_1, \ldots, \delta_p$ définis sur les domaines $\mathcal{B}_j$ en une dissimilarité globale sur E :

$$d : E \times E \to \mathbb{R}^+$$

$$(x_k, x_l) \rightsquigarrow d(x_k, x_l) = \left(\sum_{j=1}^{p} \delta_j^2(x_{kj}, x_{lj})\right)^2$$

où δ_j est une des dissimilarités définies précédemment.

2.7.3 Le cas des variables modales

Considérons $E = \{x_1, \ldots, x_n\}$ un ensemble de n objets décrits par p variables multivaluées $Y_1, \ldots, Y_p$ dont les espaces d'observations sont respectivement $\mathcal{Y}_1, \ldots, \mathcal{Y}_p$.
Le principe est similaire que celui du cas précédent.
Nous notons m_j le nombre de modalités que peut prendre la variable Y_j. Chaque objet symbolique $x_k \in E$ peut dès lors être représenté par un vecteur de dimension $m_1 + \ldots + m_p$, c'est-à-dire

$$x_k = ((q_{1,x_k}(c_1), \ldots, q_{1,x_k}(c_{m_1})), \ldots, (q_{p,x_k}(c_1), \ldots, q_{p,x_k}(c_{m_p})).$$

où $q_{j,x_k}(c_s)$ représente la valeur de distribution $\pi_{j,k}$ associée à la catégorie $c_s (s = 1, \ldots, m_j)$ de Y_j.

De cette manière, nous pouvons obtenir la même matrice $\widetilde{X}$ que celle définie pour les variables multivaluées catégoriques et nous pouvons définir les trois mêmes distances.

d'individus.
La distance la plus souvent utilisée est la distance euclidienne :

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{p}(x_{il} - x_{jl})^2} \quad \forall x_i, x_j \in E$$

3.2 Structures classificatoires

3.2.1 Les partitions

Une **partition** d'un ensemble E est un ensemble de parties non vides $P = (C_1, \ldots, C_k)$, d'intersections vides deux à deux et dont la réunion forme Ω, c'est-à-dire :

1. $\forall l \in \{1, \ldots, k\}, C_l \neq \emptyset$;
2. $\forall l, m \in \{1, \ldots, k\}, l \neq m, C_l \bigcap C_m = \emptyset$;
3. $\bigcup_{i=1}^{k} = E$.

Exemple 3.2.1 *Supposons un ensemble de onze points décrits par deux variables. Graphiquement, une partition de ces points en trois classes peut être donnée par :*

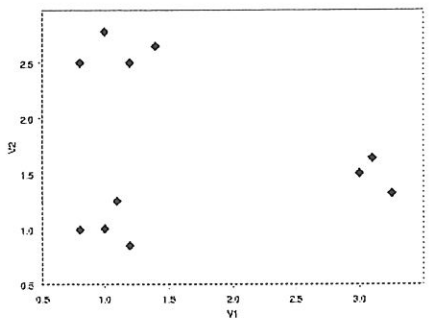


FIG. 3.1 – Représentation graphique d'une partition

3.2.2 Les hiérarchies

Une hiérarchie permet de représenter l'ensemble E des individus par un ensemble de partitions emboîtées.

Soit E un ensemble fini, H un ensemble de parties (appelées paliers) non vides de E. H est une hiérarchie sur E si :

1. $E \in H$, c'est-à-dire que le palier le plus haut contient tous les individus ;

2. $\forall x_i \in E, \{x_i\} \in H$,c'est-à-dire que chaque singleton appartient à H;
3. $\forall h, h' \in H$, nous avons $h \bigcap h' \neq \emptyset \Rightarrow h \subset h'$ ou $h' \subset h$.

Une hiérarchie est représentée graphiquement par un arbre hiérarchique appelé aussi dendogramme.

Exemple 3.2.2 *Le dendogramme associé à l'exemple précédent est donné par :*

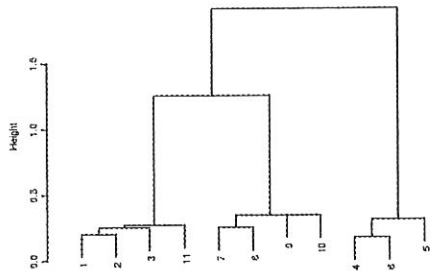


FIG. 3.2 – Représentation graphique d'une hiérarchie

3.3 Les méthodes de classification

Dans cette section, nous présentons briévement deux types de méthodes de classification :

- les méthodes hiérarchiques ;
- les méthodes non-hiérarchiques.

Celles-ci se différencient par la structure classificatoire qu'elles engendrent : des hiérarchie de partitions pour les premières, des partitions pour les secondes.

3.3.1 Les méthodes hiérarchiques agglomératives

Ces méthodes permettent de générer une suite de partitions, de moins en moins fines, de sorte que les groupements successifs d'individus forment une hiérarchie.

L'algorithme général de classification ascendante hiérarchique consiste à :

1. partir de la partition dont les classes sont réduites aux n singletons ;
2. réunir, à chaque étape, les deux individus ou groupes d'individus les plus semblables au sens du critère choisi ;
3. répéter le second point jusqu'à ce que tous les individus ne forment plus qu'une seule classe.

Le choix du critère agglomératif détermine la méthode de classification utilisée. Nous présenterons 4 méthodes hiérarchiques agglomératives ultérieurement.

3.3.2 Les méthodes non-hiérarchiques ou de partitionnement

L'idée centrale de ces méthodes est de choisir une partition initiale des individus, puis de les déplacer d'un groupe à l'autre de manière à obtenir une partition qui, au fil des itérations, améliore un critère mathématique.

Nous détaillerons une méthode non-hiérarchique qui est la méthode des nuées dynamiques.

3.4 Les méthodes hiérarchiques agglomératives

3.4.1 La méthode du saut minimum [5]

La distance entre deux classes C_i et C_j d'une partition est définie par :

$$d(C_i, C_j) = min_{x \in C_i, y \in C_j} d(x, y)$$

c'est-à-dire la plus petite distance séparant deux individus, l'un appartenant à C_i , l'autre à C_j.

Cette méthode est simple, rapide et invariante par rapport aux transformations monotones des dissimilarités. De plus, elle détecte facilement la présence de classes bien séparées et allongées, par l'effet de chaînage qu'elle engendre.

Néanmoins, cet effet de chaînage peut avoir des conséquences néfastes. En particulier si des ponts existent entre les classes. Cette méthode manque également de robustesse, dans le sens où de petites perturbations des données peuvent conduire à des partitions totalement différentes.

3.4.2 La méthode du saut maximum [5]

Dan ce cas, la distance entre deux classes C_i et C_j d'une partition est définie par :

$$d(C_i, C_j) = max_{x \in C_i, y \in C_j} d(x, y)$$

c'est-à-dire la plus grande distance séparant deux individus, l'un appartenant à C_i , l'autre à C_j.

Cette méthode à l'avantage d'être simple et invariante par rapport aux transformations monotones des dissimilarités mais elle est biaisée par rapport aux classes hypersphériques

c'est-à-dire qu'elle a tendance à retrouver des classes hypersphériques même si les données suggèrent des classes allongées.

3.4.3 La méthode du centroïde [5]

Cette méthode est basée sur la notion de centre de gravité. Celui d'un groupe l est défini par :

$$g^{(l)} = (g_1^{(l)}, \ldots, g_p^{(l)}) \ avec \ g_j^{(l)} = \frac{1}{n_l} \sum_{i=1}^{n_l} x_{ij}^{(l)}$$

où n_l est le nombre d'individus dans la classe l.

La distance entre deux groupes est définie par la distance entre les centroïdes respectifs :

$$d(C_i, C_j) = d(g^{(i)}, g^{(j)}).$$

Cette méthode consiste donc à regrouper les deux groupes dont les centroïdes sont les plus proches.
Cette méthode est biaisée par rapport aux classes hypersphériques. Par ailleurs, elle garantit que 2 groupes fusionnés sont composés d'individus en moyenne proches, mais ces 2 groupes ne sont pas pour autant homogènes. Donc, il y a perte des caractéristiques des petits groupes.

3.4.4 La méthode de Ward [6]

La méthode de Ward est l'une des méthodes les plus utilisées en classification automatique. Comme nous allons le voir, elle est basée sur le critère des moindres carrés (ou critère de la variance).

Nous allons définir les deux types d'erreur suivants :
- **l'erreur associée à chaque classe** C_l

$$\begin{aligned} e_l^2 &= \sum_{i=1}^{n_l} \sum_{j=1}^{p} (x_{ij}^{(l)} - g_j^{(l)})^2 \\ &= \sum_{i=1}^{n_l} \|x_i^{(l)} - g^{(l)}\|^2 \end{aligned}$$

où n_l est le nombre d'individus dans la classe C_l.
Cette valeur représente, pour une classe considérée, la somme des distances de chaque individu de cette classe à son centroïde.

– **l'erreur associée à une partition** :

$$E_k^2 = \sum_{l=1}^{k} e_l^2$$

qui n'est rien d'autre que le critère des moindres carrés.

Le principe de la méthode de Ward est donc de fusionner les classes pour lesquelles l'accroissement du critère des moindres carrés ΔE_{rs}^2 est minimum. ΔE_{rs}^2 représente l'accroissement de la valeur de E^2 lorsque nous regroupons les classes r et s.

Si les classes r et s sont fusionnées en la classe t, nous avons que :

$$\Delta E_{rs}^2 = e_t^2 - e_r^2 - e_s^2.$$

Le critère de la variance est optimisé à chaque étape, mais cela n'assure en rien que la partition finale en k classes est celle pour laquelle le critère de la variance est minimum.

Comme les méthodes du saut maximum et du centroïde, cette méthode est biaisée par rapport aux classes hypersphériques.

3.5 Classification de données symboliques

Nous avons défini un certain nombre de méthodes de classification automatique pour les données classiques. Il existe également des méthodes de classification pour les données symboliques, à savoir : SCLUST, la méthode divisive de Chavent, les arbres de clustering,...

Dans le cadre de ce mémoire, nous nous sommes principalement focalisés sur SCLUST qui est une extension de la méthode des nuées dynamiques aux données symboliques. Nous la présenterons au chapitre 4.

Chapitre 4

La méthode des nuées dynamiques

4.1 Notion d'inertie

4.1.1 Inertie par rapport à un point [7]

Considérons un ensemble de n individus $E = \{x_1, \ldots, x_n\}$ décrits par p variables quantitatives.

La matrice de données s'écrit :

$$\widetilde{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

Les n individus forment alors un nuage de points dans $\mathbb{R}^p$.

Définition 4.1.1
On appelle ***inertie de E par rapport à un point*** $a \in \mathbb{R}^p$ *la quantité :*

$$I_a(E) = \sum_{i=1}^{n} d^2(a, x_i)$$

où d est la distance euclidienne.

4.1.2 Théorème de Huygens[7]

Huygens a établi que le centre de gravité g est le point par rapport auquel l'inertie d'un nuage de points est minimale.

L'inertie du nuage de points de E par rapport à un point a quelconque est liée à l'inertie par rapport au centre de gravité $g \in E$ par la relation :

$$I_a(E) = I_g(E) + nd^2(g, a)$$

L'inertie $I_g(E)$, notée T, est définie comme l'inertie totale du nuage de points E. Nous allons maintenant montrer comment l'inertie peut être décomposée.

4.1.3 Inerties associées à une partition [7]

Définissons :
- $P = \{C_1, \ldots, C_k\}$ une partition de l'ensemble E en k classes ;
- $g^{(l)} = \frac{1}{n_l} \sum_{i=1}^{n_l} x_i^{(l)}$ le centre de gravité de la classe l.

A cette partition P sont associées trois inerties :

- **l'inertie totale** T, définie par :

$$T = I_g(E) = \sum_{i=1}^{n} d^2(x_i, g).$$

  Il s'agit de la somme des carrés des distances de chaque point x_i au centre de gravité global g.

- **l'inertie intra-classes** W :

$$W = \sum_{l=1}^{k} \sum_{x_i \in C_l} d^2(x_i, g^{(l)})$$

  représente la somme sur chaque classe des inerties de C_l à son centre de gravité $g^{(l)}$.

- **l'inertie inter-classes** B :

$$B = \sum_{l=1}^{k} n_l d^2(g^{(l)}, g).$$

  est l'inertie du nuage formé des k centres de gravité $g^{(l)}$ locaux, pondérée par le nombre d'individus de la classe correspondante, par rapport au centre de gravité global g.

Les trois quantités que nous venons de décrire vérifient la relation fondamentale suivante :

$$T = W + B.$$

Notons que l'inertie totale T est indépendante de la partition et donc plus l'inertie inter-classes B est grande, plus l'inertie intra-classes W est petite.

4.1.4 Décomposition des inerties sur les classes et les variables [7]

Pour pouvoir analyser une partition, il est utile de décomposer les inerties associées suivant les classes et les variables.
L'inertie totale T peut s'écrire :

$$T = \sum_{\ell=1}^{k} \sum_{j=1}^{p} T_j^{(\ell)} \quad avec \quad T_j^{(\ell)} = \sum_{x_i \in C_\ell} (x_{ij} - g_j)^2.$$

$T_j^{(\ell)}$ est l'écart, pour la variable j, des points de la classe C_ℓ au centre de gravité global g_j.

L'inertie intra-classes W peut être définie par :

$$W = \sum_{\ell=1}^{k} \sum_{j=1}^{p} W_j^{(\ell)} \quad avec \quad W_j^{(\ell)} = \sum_{x_i \in C_\ell} (x_{ij} - g_j^{(\ell)})^2.$$

$W_j^{(\ell)}$ est l'inertie de la classe C_ℓ, pour la variable j, par rapport au centre de gravité local $g_j^{(\ell)}$ de cette même classe.

L'inertie inter-classes B peut être définie par :

$$B = \sum_{\ell=1}^{k} \sum_{j=1}^{p} B_j^{(\ell)} \quad avec \quad B_j^{(\ell)} = (g_j^{(\ell)} - g_j)^2.$$

$B_j^{(\ell)}$ est l'écart, pour la variable j, du centre de gravité local $g_j^{(\ell)}$ de la classe C_ℓ au centre de gravité global g_j.

De plus, les différentes inerties peuvent se décomposer de manière additive sur les variables et sur les classes.

Si nous notons $\forall j \in \{1, \ldots, p\}$:

$$T_j = \sum_{l=1}^{k} T_j^{(\ell)} \qquad W_j = \sum_{\ell=1}^{k} W_j^{(\ell)} \qquad B_j = \sum_{\ell=1}^{k} B_j^{(\ell)}$$

nous avons :

$$T = \sum_{j=1}^{p} T_j \qquad W = \sum_{j=1}^{p} W_j \qquad B = \sum_{j=1}^{p} B_j$$

où T_j, W_j, B_j expriment la contribution de la variable j pour les inerties concernées.

Si nous notons $\forall \ell \in \{1, \ldots, k\}$:

$$T^l = \sum_{j=1}^{p} T_j^{(\ell)} \qquad W^\ell = \sum_{j=1}^{p} W_j^{(\ell)} \qquad B^l = \sum_{j=1}^{p} B_j^{(\ell)}$$

nous avons :

$$T = \sum_{\ell=1}^{k} T^{(\ell)} \qquad W = \sum_{\ell=1}^{k} W^{(\ell)} \qquad B = \sum_{\ell=1}^{k} B^{(\ell)}$$

où $T^{(\ell)}, W^{(\ell)}, B^{(\ell)}$ expriment la contribution de la classe ℓ pour les inerties concernées.

4.2 La méthode des nuées dynamiques dans le cas de données classiques [7]

4.2.1 Introduction

La méthode des nuées dynamiques cherche à optimiser un critère qui exprime l'adéquation entre une classification des objets et un mode de représentation des classes de cette classification. Le problème d'optimisation associé à cette méthode se pose alors en termes de recherche simultanée de la classification et de la représentation des classes de cette classification parmi un ensemble de classifications et de représentations possibles, qui optimise le critère.

Nous commencerons par définir le principe général de l'algorithme avant de formaliser les principales étapes.

4.2.2 Principe général

L'algorithme des nuées dynamiques nécessite tout d'abord la définition d'un mode de représentation de toute classe d'individus.
Cette représentation , ou **noyau**, peut être, par exemple :

- une droite ;
- un ensemble de points ;
- un centre de gravité ;
- le point le plus proche du centre de gravité.

Le déroulement de l'algorithme est le suivant :

- Dans un premier temps, k noyaux estimés ou tirés au hasard sont choisis parmi une famille de noyaux admissibles, appelée **espace de représentation**, noté $\mathcal{L}$.
- Ensuite, les points sont affectés au noyau le plus proche. Nous obtenons alors une partition en k classes dont les nouveaux noyaux sont calculés.

– Le procédé est alors recommencé avec les nouveaux noyaux.

Sous certaines conditions portant sur les fonctions qui permettent d'affecter les points aux classes et de calculer les noyaux, cet algorithme fait décroître un critère W qui mesure l'adéquation entre les classes et leur noyau respectif, c'est-à-dire la ressemblance des noyaux à leur classe.

Le critère s'exprime de la manière suivante :

$$W : \mathcal{P}_k \times \mathcal{L}_k \to \mathbb{R}^+$$

$$(P, L) \rightsquigarrow W(P, L) = \sum_{\ell=1}^{k} D(L_\ell, C_\ell)$$

- $\mathcal{P}_k$ est l'ensemble des partitions $P = (C_1, \ldots, C_k)$ en k classes de E ;
- $\mathcal{L}_k = E^k$ est l'ensemble des n-uplets de noyaux $L = (L_1, \ldots, L_k)$ avec $L_\ell \in E$ dans le cas où chaque noyau est un point de E ; nous avons $\mathcal{L}_k = \mathbb{R}^k$ si les k noyaux sont les centres de gravité des k classes.
- D est une mesure d'adéquation du noyau L_ℓ à la classe C_{ell}, c'est-à-dire qu'une petite valeur de D exprime une bonne adéquation entre L_ℓ et C_ℓ.

Ainsi, à chaque itération de l'algorithme, la diminution de la valeur du critère exprime une augmentation globale de l'adéquation entre les classes et les noyaux.

4.2.3 Aspect formel

L'objectif de l'algorithme des nuées dynamiques est de déterminer un couple $(P^*, L^*) \in \mathcal{P}_k \times \mathcal{L}_k$ où :

- $\mathcal{P}_k$ est l'ensemble des partitions en k classes ;
- $\mathcal{L}_k$ est un espace de représentation

qui minimise un critère mathématique

$$W : \mathcal{P}_k \times \mathcal{L}_k \to \mathbb{R}^+$$

c'est-à-dire que $W(P^*, L^*) = \min\{W(P, L) | P \in \mathcal{P}_k \ et \ L \in \mathcal{L}_k\}$.

Ce critère peut être minimisé par l'utilisation successive d'une étape dite de représentation et d'une étape dite d'affectation, et ce itérativement jusqu'à obtenir la convergence.

La méthode des nuées dynamiques consiste à :

1. Choisir un espace de représentation $\mathcal{L}_k$;

2. Définir un critère $W : \mathcal{P}_k \times \mathcal{L}_k \to \mathbb{R}^+$ qui permet de mesurer l'adéquation entre toute partition $P \in \mathcal{P}_k$ et toute représentation $L \in \mathcal{L}_k$ de cette partition ;
3. Chercher simultanément la partition$P \in \mathcal{P}_k$ et une représentation L de cette partition de sorte que P et L aient la meilleure adéquation au sens du critère W.
4. Ce problème peut être résolu grâce à l'algorithme des nuées dynamiques. Celui-ci consiste à utiliser itérativement :
   - une fonction de représentation $g : \mathcal{P}_k \to \mathcal{L}_k$ qui permet de calculer k noyaux à partir de la partition en k classes et
   - une fonction d'affectation $f : \mathcal{L}_k \to \mathcal{P}_k$ qui construit la partition en k classes en affectant chaque individu au noyau dont il est le plus proche.

   L'initialisation est effectuée à l'aide d'une partition $P^{(0)} \in \mathcal{P}_k$ ou d'une représentation $L^{(0)} \in \mathcal{L}_k$ estimées ou tirées au hasard.
5. Nous pouvons alors définir une suite $u_n = W(v_n)$ avec $v_n = (P^n, L^n)$ où :
   - $P^n \in \mathcal{P}_k$ est la partition en k classes obtenues à l'itération n ;
   - $L^n \in \mathcal{L}_k$ est la représentation de P^n obtenue à l'aide de la fonction g.

   Si les fonctions d'affectation f, permettant de passer de L^n à $P^{(n+1)}$, et de représentation g, permettant de passer de $P^{(n+1)}$ à $L^{(n+1)}$, sont bien choisies, nous pouvons montrer que les suites u_n et v_n sont convergentes et que la suite u_n est décroissante, ce qui signifie que l'algorithme fait bien décroître la valeur du critère à chaque itération jusqu'à stabilisation.

4.2.4 Le cas du centre de gravité

Cette section traite le cas où les noyaux ou prototypes de chaque classe sont définis par les centres de gravité respectifs.

Espace de représentation

L'espace des individus est l'espace $\mathbb{R}^p$, de même que l'espace de représentation $\mathcal{L}$ d'une classe. La mesure d'adéquation est définie par :

$$D : \mathcal{P} \times \mathcal{B} \to \mathbb{R}^+$$

tel que

$$\forall A \in \mathcal{P} \quad et \quad \forall x \in \mathbb{R}^p \quad D(A, x) = \sum_{a \in A} d^2(a, x) = I_x(A)$$

où d est une distance euclidienne.

Cette mesure d'adéquation D n'est rien d'autre que l'inertie de l'ensemble A par rapport au point x.

La fonction de représentation g

Le point x qui minimise l'inertie de A par rapport à x est, par le théorème de Huygens rappelé au début de ce chapitre, le centre de gravité de A.

La fonction de représentation g qui, à toute partition $P = (C_1, \ldots, C_k)$ associe sa représentation $L = (L_1, \ldots, L_k)$ est donc définie par :

$$\begin{aligned} g : \quad \mathcal{P}_k &\rightarrow \quad \mathcal{L}_k \\ (C_1, \ldots, C_k) &\rightsquigarrow (g^{(1)}, \ldots, g^{(k)}) \end{aligned}$$

où $g^{(\ell)}$ est le centre de gravité de la classe ℓ.

La fonction d'affectation f

La fonction d'affectation f est définie par :

$$\begin{aligned} f : \quad \mathcal{L}_k &\rightarrow \quad \mathcal{P}_k \\ (g^{(1)}, \ldots, g^{(k)}) &\rightsquigarrow (C_1, \ldots, C_k) \end{aligned}$$

où $C_\ell = \{x \in \mathbb{R}^p | d(x, g^{(\ell)}) \leq d(x, g^{(m)}) \ \forall m \in \{1, \ldots, k\} \ \text{ et } \ \ell < m \text{ en cas d'égalité } \}$

c'est-à-dire que C_ℓ est la classe composée des individus les plus proches, au sens de la métrique choisie, de son centre de gravité $g^{(\ell)}$.

Le problème d'optimisation

Il s'agit de chercher le couple $(P^*, L^*) \in \mathcal{P}_k \times \mathcal{L}_k$ minimisant le critère d'adéquation W entre la partition $P = (C_1, \ldots, C_k)$ et sa représentation $L = (g^{(1)}, \ldots, g^{(k)})$ défini par :

$$W(P, L) = \sum_{\ell=1}^{k} D(C_\ell, g^{(\ell)}) = \sum_{\ell=1}^{k} \sum_{x_i \in C_\ell} d^2(x_i, g^{(\ell)})$$

où $g^{(\ell)}$ est le centre de gravité de la classe ℓ.

Nous pouvons également écrire ce critère sous la forme :

$$W(P, L) = \sum_{\ell=1}^{k} W_\ell$$

où W_ℓ est l'inertie de la classe C_ℓ par rapport à son centre de gravité $g^{(\ell)}$.

Le critère $W(P, L)$ est alors l'inertie intra-classes de la partition P.

Donc, lorsque les noyaux sont les k centres de gravité, la méthode cherche à minimiser l'inertie intra-classes W et par conséquent à minimiser l'inertie inter-classes B en vertu de la relation

$$T = W + B$$

Algorithme

En utilisant les fonctions f et g de façon itérative, nous définissons une suite $v_n = (P^n, L^n)$ et une suite $u_n = W(v_n)$.

Nous avons alors les deux propositions suivantes dont les démonstrations se trouvent dans [7].

Proposition 1

La suite u_n converge en décroissant.

Proposition 2

La suite v_n est stationnaire.

4.3 La méthode des nuées dynamiques dans le cas de données symboliques

4.3.1 Le cas des données de type intervalle

Supposons un ensemble $E = \{x_1, \ldots, x_n\}$ de n objets symboliques décrits par p variables de type intervalle. Ainsi, un objet symbolique x_k peut-être représenté par un hyperrectangle dans un espace euclidien p-dimensionnel, c'est-à-dire :

$$x_k = ([\alpha_{k1}, \beta_{k1}], \ldots, [\alpha_{kp}, \beta_{kp}])$$

Le cas qui nous intéresse est bien évidemment celui traité par le programme SCLUST, dans lequel les noyaux ou prototypes de chaque classe sont définis par les hyperrectangles de gravité respectifs. Ceci est une extension du centre de gravité vu dans la section précédente pour les données symboliques.

L'hyperrectangle de gravité d'un ensemble de n objets symboliques décrits par p variables de type intervalle est défini par :

$$\left(\left[\frac{1}{n}\sum_{i=1}^{n}\alpha_{i1}, \frac{1}{n}\sum_{i=1}^{n}\beta_{i1}\right], \left[\frac{1}{n}\sum_{i=1}^{n}\alpha_{i2}, \frac{1}{n}\sum_{i=1}^{n}\beta_{i2}\right], \ldots, \left[\frac{1}{n}\sum_{i=1}^{n}\alpha_{ip}, \frac{1}{n}\sum_{i=1}^{n}\beta_{ip}\right]\right)$$

Espace de représentation

L'espace des objets de l'espace $\mathcal{I}^p$ des hyperrectangles fermés bornés, de même que l'espace de représentation $\mathcal{L}$ d'une classe. La mesure d'adéquation est définie par :

$$D : \mathcal{P}(E) \times \mathcal{B} \to \mathbb{R}^+$$

tel que

$$\forall A \in \mathcal{P}(E) \quad et \quad \forall x_i = ([\alpha_{i1}, \beta_{i1}], \ldots, [\alpha_{ip}, \beta_{ip}]) \in \mathcal{I}^p \qquad D(A, x_i) = \sum_{a \in A} d^2(a, x_i)$$

où d est une mesure de dissimilarité. Pour rappel, trois distances ont été définies pour les objets symboliques de type intervalle dans le chapitre 2.

La fonction de représentation g

La fonction de représentation g qui, à toute partition $P = (C_1, \ldots, C_k)$ associe sa représentation $L = (L_1, \ldots, L_k)$ est donc définie par :

$$\begin{array}{rccc} g : & \mathcal{P}_k & \to & \mathcal{L}_k \\ & (C_1, \ldots, C_k) & \rightsquigarrow & (g^{(1)}, \ldots, g^{(k)}) \end{array}$$

où

$$g^{(\ell)} = \left(\left[\frac{1}{n_\ell} \sum_{x_i \in C_\ell} \alpha_{i1}, \frac{1}{n_\ell} \sum_{x_i \in C_\ell} \beta_{i1} \right], \ldots, \left[\frac{1}{n_\ell} \sum_{x_i \in C_\ell} \alpha_{ip}, \frac{1}{n_\ell} \sum_{x_i \in C_\ell} \beta_{ip} \right] \right)$$

est le prototype de la classe ℓ.

La fonction d'affectation f

La fonction d'affectation f est définie par :

$$\begin{array}{rccc} f : & \mathcal{L}_k & \to & \mathcal{P}_k \\ & (g^{(1)}, \ldots, g^{(k)}) & \rightsquigarrow & (C_1, \ldots, C_k) \end{array}$$

où $C_\ell = \{x \in E | d(x, g^{(\ell)}) \leq d(x, g^{(m)}) \forall m \in \{1, \ldots, k\}$ et $\ell < m$ en cas d'égalité $\}$

c'est-à-dire que C_{ell} est la classe composée des individus les plus proches, au sens de la métrique choisie, de son hyperrectangle de gravité $g^{(\ell)}$.

Le problème d'optimisation

Il s'agit de chercher le couple $(P^*, L^*) \in \mathcal{P}_k \times \mathcal{L}_k$ minimisant le critère d'adéquation W entre la partition $P = (C_1, \ldots, C_k)$ et sa représentation $L = (g^{(1)}, \ldots, g^{(k)})$ défini par :

$$W(P, L) = \sum_{\ell=1}^{k} D(C_\ell, g^{(\ell)}) = \sum_{\ell=1}^{k} \sum_{x_i \in C_\ell} d^2(x_i, g^{(\ell)})$$

où $g^{(\ell)}$ est l'hyperrectangle de gravité de la classe ℓ.

Nous pouvons également écrire ce critère sous la forme :

$$W(P, L) = \sum_{\ell=1}^{k} W_\ell$$

où W_ℓ est l'inertie de la classe C_ℓ par rapport à son noyau $g^{(\ell)}$.

Le critère $W(P, L)$ est alors l'inertie intra-classes de la partition P.

Donc, lorsque les noyaux sont les k hyperrectangles de gravité, la méthode cherche à minimiser l'inertie intra-classes W et en vertu de la relation $T = W + B$, cela revient à minimiser l'inertie inter-classes B.

4.3.2 Le cas des variables multivaluées et modales

Considérons un ensemble $E = \{x_1, \ldots, x_n\}$ de n objets symboliques décrits par p variables multivaluées ou modales $Y_1, \ldots, Y_p$ où Y_j a pour domaine $\mathcal{Y}_j$. Ainsi, un objet symbolique x_k peut-être représenté par un vecteur à $m_1 + \ldots + m_p$ dimensions, c'est-à-dire :

$$x_k = ((q_{1,x_k}(c_1), \ldots, q_{1,x_k}(c_{m_1})), \ldots, (q_{p,x_k}(c_1), \ldots, q_{p,x_k}(c_{m_p})).$$

où

- m_j représente le nombre de modalités que peut prendre la variable Y_j et
- $q_{j,x_k}(C_\ell)$ est la valeur de distribution $\pi_{j,k}$ associée à la catégorie $c_\ell(\ell = 1, \ldots, m_j)$ de Y_j.

Comme pour les variables de type intervalle, nous devons construire des prototypes qui résument toute l'information contenue dans les différentes classes sur les objets symboliques.

Le prototype $g^{(\ell)}$ de la classe C_ℓ est construit de la manière suivante :

$$\left(\left(\frac{1}{n_\ell}\sum_{x_i \in C_\ell} q_{1,x_i}(c_1), \ldots, \frac{1}{n_\ell}\sum_{x_i \in C_\ell} q_{1,x_i}(c_{m_1})\right), \ldots, \left(\frac{1}{n_\ell}\sum_{x_i \in C_{ell}} q_{p,x_i}(c_1), \ldots, \frac{1}{n_\ell}\sum_{x_i \in C_\ell} q_{p,x_i}(c_{m_p})\right)\right)$$

où n_ℓ est le nombre d'individus dans la classe C_ℓ.

La mesure d'adéquation D, la fonction de représentation g, la fonction d'affectation f et le problème d'optimisation sont définis comme dans le cas précédent.

Chapitre 5

Méthodes de détermination du nombre de classes

5.1 Introduction

Les méthodes de classification hiérarchiques ou non-hiérachiques que nous venons de détailler fournissent, pour les unes, des hiérarchies de partitions et pour les autres, des partitions. Néanmoins, la structure classificatoire obtenue suppose un nombre de classes fixé a priori. Mais combien y a-t-il réellement de classes présentes dans les données ? Nous ne le savons pas au juste. Les méthodes de détermination du nombre de classes permettent de répondre à cette question.

5.2 Les méthodes de Milligan et Cooper

En 1985, Milligan et Cooper [4] ont comparé trente méthodes de déterminantion du nombre de classes. D'après leur étude, ils ont réalisé un classement des différentes méthodes. Nous présenterons les cinq "meilleures" méthodes issues de leur classement.

5.2.1 La méthode de Calinski et Harabasz (1974) [8]

Considérons la matrice de données classiques

$$\widetilde{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix}$$

Définissons :
- la matrice de **dispersion totale** :

$$M_T = \sum_{i=1}^{n} (x_i - g)(x_i - g)'$$

où
- x_i est le i$^{\text{ème}}$ individu ;
- n est le nombre d'individus ;
- g est le centre de gravité du nuage de points.

- la matrice de **dispersion intra-classes** :

$$M_W = \sum_{\ell=1}^{k} M_W^{(\ell)} = \sum_{\ell=1}^{k} \sum_{i=1}^{n_\ell} (x_i^{(\ell)} - g^{(\ell)})(x_i^{(\ell)} - g^{(\ell)})'$$

où
- k est le nombre de classes ;
- n_ℓ est le nombre d'individus de la $\ell^{\text{ème}}$ classe ;
- $x_i^{(\ell)}$ est le i$^{\text{ème}}$ individu de la $\ell^{\text{ème}}$ classe ;
- $g^{(\ell)} = \frac{1}{n_{\ell ell}} \sum_{i=1}^{n_\ell} x_i^{(\ell)}$ est le centre de gravité de la $\ell^{\text{ème}}$ classe.

- la matrice de dispersion **inter-classes** :

$$M_B = \sum_{\ell=1}^{k} \sum_{i=1}^{n_\ell} (g^{(\ell)} - g)(g^{(\ell)} - g)'$$

- k est le nombre de classes ;
- n_ℓ est le nombre d'individus de la $\ell^{\text{ème}}$ classe ;
- g est le centre de gravité du nuage de points ;
- $g^{(\ell)}$ est le centre de gravité de la $\ell^{\text{ème}}$ classe.

Ces matrices vérifient :

$$M_T = M_B + M_W$$

L'indice permettant de définir le nombre de classes est :

$$CH = \frac{tr(M_B)/(k-1)}{tr(M_W)/(n-k)}$$

Il faut choisir le nombre de classes k tel que CH a un maximum relatif ou absolu ou bien un écart important entre deux de ses valeurs consécutives.

5.2.2 La méthode de Duda et Hart [9]

Soit un ensemble de n individus sur lesquels p variables ont été observées.

On définit

$$J_e(k) = \sum_{\ell=1}^{k} \sum_{i=1}^{n_\ell} \sum_{j=1}^{p} (x_{ij}^{(\ell)} - g_j^{(\ell)})^2$$

qui est la somme des carrés des erreurs qui surviennent si les n individus sont représentés par les k centres des classes.

Le critère de détermination du nombre de classes est basé sur le test d'hypothèse suivant :

$$\begin{cases} H_0 : \text{Les points sont issus d'une même population Normale de moyenne } \mu \text{ et de} \\ \quad \text{matrice de covariance } \sigma^2 I. \\ H_1 : \text{Les points sont issus de deux populations Normales.} \end{cases}$$

Nous rejettons H_0 au niveau de signification α, si

$$\frac{J_e(2)}{J_e(1)} < 1 - \frac{2}{\pi p} - z_{1-\alpha}\sqrt{\frac{2(1-\frac{8}{\pi^2 p})}{np}}$$

où
- n est le nombre d'individus ;
- p est le nombre de variables ;
- $z_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi Normale.

Ce test est effectué à chaque niveau de la hiérarchie, entre 2 groupes candidats à la fusion. A chaque niveau de la hiérarchie, on calcule

$$\frac{-\frac{J_e(2)}{J_e(1)} + 1 - \frac{2}{\pi p}}{\sqrt{\frac{2(1-\frac{8}{\pi^2 p})}{np}}} \tag{5.1}$$

La règle de décision est la suivante : on rejette H_0 qui correspond à l'existence d'un seul groupe si la valeur de (5.1) dépasse la valeur choisie pour $z_{1-\alpha}$.
Dès lors, si k_0 est la première valeur pour laquelle on rejette H_0, on conclut que $(k_0 + 1)$ classes sont présentes dans les données.

Différentes valeurs ont été proposées pour $z_{1-\alpha}$. Milligan et Cooper ont suggéré $z_{1-\alpha} = 3.20$; alors que Gordon obtient de meilleurs résultats avec un choix de $z_{1-\alpha}$ valant 4.

5.2.3 La méthode du C-index [10]

Soit $E = \{x_1, \ldots, x_n\}$. Nous définissons

$$C(x_i, x_j) = \begin{cases} 1 & si\ x_i \text{ et } x_j \text{ sont dans la même classe} \\ 0 & \text{sinon} \end{cases}$$

et

$$\begin{aligned} \Gamma &= \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} C(x_i, x_j) \\ &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij} C(x_i, x_j) \\ &= \sum_{l=1}^{k} \sum_{\substack{i,j=1 \\ i<j}}^{n_\ell} d(x_i^{(\ell)}, x_j^{(\ell)}) \end{aligned}$$

où
- d est la distance euclidienne ;
- k est le nombre de classes ;
- n_ℓ est le nombre d'individus dans la classe ℓ.

Nous normalisons Γ afin de déterminer le nombre de classes. La normalisation a été proposée par Dalrymphe-Arford et permet de définir l'indice

$$indexC = \frac{\Gamma - min\Gamma}{max\Gamma - min\Gamma}$$

Si la partition contient r dissimilarités entre paires d'individus :
- $min\Gamma$ est la somme des r plus petites dissimilarités entre paires d'individus ;
- $max\Gamma$ est la somme des r plus grandes dissimilarités entre paires d'individus.

Le nombre de classes est indiquée par la valeur minimale de l'index C. La valeur idéale vaut 0.

5.2.4 La méthode Gamma [11]

Considérons n objets $x_1, \ldots, x_n$ répartis en k classes tel que $d(x_i, x_j)$ est la distance euclidienne entre x_i et x_j.

Définissons

$$T_\ell(x_i, x_j) = \begin{cases} 0 & si\ x_i \text{ et } x_j \text{ sont dans la même classe} \\ 1 & \text{sinon} \end{cases}$$

Si $T_\ell(x_i, x_j) = 0$, ce qui signifie que x_i et x_j appartiennent à la même classe, nous posons

$$n_\ell(x_i, x_j) = \mid \{\{x_r, x_s\} : T_\ell(x_r, x_s) = 1 \ et \ d(x_r, x_s) < d(x_i, x_j)\} \mid$$

qui est le nombre de couples d'objets qui n'appartiennent pas à la même classe et qui sont plus proches que x_i et x_j.

L'indice α_ℓ est défini par

$$\alpha_\ell = \frac{\sum_{i<j} n_\ell(x_i, x_j)}{max \sum_{i<j} n_\ell(x_i, x_j)}$$

où

- le maximum est pris sur toutes les partitions possibles en k classes en gardant le même nombre d'objets par classe,
- les sommes sont réalisées sur les paires d'objets $\{x_i, x_j\}$ qui appartiennet à la même classe.

L'indice α_ℓ représente "une" proportion de paires d'objets qui sont dans de "mauvais groupes" ; dans le sens où deux éléments d'une même classe sont plus proches que deux objets de classes différentes.

L'indice γ s'exprime alors

$$\gamma = 1 - 2\alpha_\ell.$$

Cet indice prend des valeurs comprises entre -1 et 1 étant donné que α_l est compris entre 0 et 1.
Si tous les objets sont placés correctement alors $\sum_{i<j} n_l(x_i, x_j) = 0$ et par conséquent $\alpha_l = 0$ et $\gamma = 1$.
Si par contre un maximum d'objets est mal placés alors $\alpha_\ell = 1$ et $\gamma = -1$.
Nous cherchons donc la valeur maximale de l'indice γ, celle-ci devant être idéalement 1.

5.2.5 La méthode de Beale [12]

Comme pour la méthode de Duda et Hart, celle de Beale permet de décider si une fusion entre deux groupes est justifiée ou non et est donc basée sur un test d'hypothèse défini sur le nombre de classes et s'exprimant :

$$\begin{cases} H_0 : & t = 1 \\ H_1 : & t = 2 \end{cases}$$

Soit un ensemble de n individus sur lesquels p variables ont été mesurées.
Définissons

$$W_1 = \sum_{i=1}^{n^*} \sum_{j=1}^{p} (x_{ij} - g_j)^2 = I(C_1 \cup C_2)$$

$$W_2 = \sum_{\ell=1}^{2} \sum_{i=1}^{n_\ell} \sum_{j=1}^{p} (x_{ij}^{(\ell)} - g_j^{(\ell)})^2 = I(C_1) + I(C_2)$$

où

- C_1 et C_2 sont les deux classes candidates à la fusion,
- n_ℓ est le nombre d'éléments dans la classe C_ℓ

– n^* est le nombre d'éléments dans $C_1 \cup C_2$

La statistique du test est

$$W = \frac{\frac{W_1 - W_2}{W_2}}{\left(\frac{n-1}{n-2}\right) 2^{\frac{2}{p}-1}}$$

qui est à comparer avec une distribution de Fisher-Snedecor à p degrés de liberté au numérateur et $(n-2)p$ degré de liberté au dénominateur.

La règle de décision est le suivante :
On rejette H_0 au niveau de signification α, si $W > F_{p,(n-2)p,1-\alpha}$, c'est-à-dire si $W > 5.3$ avec $\alpha = 0.005$ ou si $W > 4.61$ si $\alpha = 0.01$.

Si k_0 est la première valeur qui conduit au rejet de la fusion de deux groupes, le test de Beale indique que le bon nombre de classes est $k_0 + 1$.

Chapitre 6

Le module DISS

Le module DISS qui a été développé dans le cadre du projet ASSO et a été élaboré par l'université de Bari. Il a pour but d'évaluer le degré de dissimilarité entre un ensemble d'objets symboliques en mémoire dans les fichiers ASSO. Il permet surtout de calculer une matrice de dissimilarité pour des objets symboliques décrits par des variables symboliques de tout type. Comme nous le savons, un objet symbolique est décrit par une collection de variables non modales et/ou modales. Le module DISS distingue deux types d'objets symboliques et utilise des mesures de dissimilarité en fonction du type d'objets rencontrés.

Nous étudierons dans ce chapitre les deux types d'objets symboliques et développerons les mesures de dissimilarité associées à ces objets. Nous expliquerons enfin comment utiliser le module DISS dans le logiciel SODAS.

6.1 Une nouvelle subdivision des objets symboliques

Dans la suite de ce mémoire, nous différencierons deux types d'objets symboliques :

1. les objets symboliques booléens (BSO) et
2. les objets symboliques probabilistiques (probabilistic) (PSO).

C'est, en effet, en fonction de ce type d'objets que nous établirons les mesures et donc la matrice de dissimilarité.

Les objets symboliques booléens sont uniquement décrits par des variables :

- catégoriques univaluées. Par exemple :

$$\text{ville}(x_k) = \text{Londres},$$

- catégoriques multivaluées. Par exemple :

$$\text{ville}(x_k) = \{\text{Londres, Paris, Rome}\},$$

– quantitatives univaluées. Par exmple :

$$\text{hauteur}(x_k) = 3.5 \text{ et}$$

– de type intervalle. Par exemple :

$$\text{hauteur}(x_k) = [3, 7].$$

Les objets symboliques probabilistiques sont des objets décrits par des variables modales uniquement.

Par exemple,

$$\text{ville}(x_k) = \{\text{Londres}(0.2), \text{Paris}(0.7), \text{Rome}(0.1)\}.$$

6.2 Mesures de dissimilarité entre objets booléens

Dans la littérature, plusieurs mesures de dissimilarité ont été proposées pour les objets symboliques booléens. Les dissimilarités implémentées dans le module DISS sont U-1, U-2, U-3, U-4, SO-1, SO-2, SO-3, SO-4, SO-5, SO-6, C-1. Sans plus attendre, nous allons détailler ces mesures.

6.2.1 La mesure de dissimilarité de Gowda et Diday (U-1)

En 1991, Gowda et Diday [3] ont proposé une mesure de dissimilarité $D(a, b)$ pour deux objets symboliques :

$$a = [Y_1 \in A_1] \wedge [Y_2 \in A_2] \wedge \ldots \wedge [Y_p \in A_p] \tag{6.1}$$

$$b = [Y_1 \in B_1] \wedge [Y_2 \in B_2] \wedge \ldots \wedge [Y_p \in B_p] \tag{6.2}$$

où chaque variable Y_j prend des valeurs dans un domaine $\mathcal{Y}_j$ et $A_j, B_j \subset \mathcal{Y}_j$.

Cette fonction de distance est donnée sous forme additive, ie

$$D(a, b) = \sum_{j=1}^{p} D(A_j, B_j) \tag{6.3}$$

Pour la variable j afin de calculer $D(A_j, B_j)$, il faut considérer trois types de mesures de dissimilarité définies pour une paire de sous-ensemble A_j, B_j et incorporant différents aspects de similarité, à savoir :

1. $D_\pi(A_j, B_j)$ due à la position,
2. $D_s(A_j, B_j)$ due à l'étendue et
3. $D_c(A_j, B_j)$ due au contenu ;

où les composantes D_π, D_s et D_c sont définies de sorte que leurs valeurs soient comprises entre 0 et 1. Toutes les composantes ne seront pas nécessairement considérées dans la définition de la distance $D(A_j, B_j)$; en effet, le choix dépendra du type de la variable Y_j.

Variables quantitatives

Par la suite, nous considèrerons des variables réelles (continues) et entières, c'est-à-dire les cas $\mathcal{Y}_j = \mathbb{R}$ ou $\mathbb{Z}$. Alors la dissimilarité entre A_j et B_j est définie par

$$D(A_j, B_j) = D_\pi(A_j, B_j) + D_s(A_j, B_j) + D_c(A_j, B_j) \tag{6.4}$$

avec les spécifications suivantes.

◇ La **composante** D_π indique les positions relatives des deux variables sur la droite réelle. Ecrivons $A_j = [\underline{a_j}, \overline{a_j}]$, $B_j = [\underline{b_j}, \overline{b_j}]$, alors D_π est défini comme suit

$$D_\pi(A_j, B_j) = \frac{|\underline{a_j} - \underline{b_j}|}{|\mathcal{Y}_j|} \tag{6.5}$$

où $|\mathcal{Y}_j|$ dénote la longueur de l'intervalle maximum de la j$^{\text{ème}}$ variable.

◇ La **composante** D_s, due à l'étendue, indique les tailles relatives des valeurs de la variable sans tenir compte de la partie commune entre elles. Elle est définie comme suit :

$$D_s(A_j, B_j) = \frac{|l_a - l_b|}{l_s} \tag{6.6}$$

où

$l_a =$ $|\overline{a_j} - \underline{a_j}|$,
$l_b =$ $|\overline{b_j} - \underline{b_j}|$,
$l_s =$ longueur d'étendue de A_j et $B_j =$ $|\max(\overline{a_j}, \overline{b_j}) - \min(\underline{a_j}, \underline{b_j})|$.

◇ Finalement, la **composante** D_c, due au contenu, est une mesure de la partie non commune entre deux valeurs d'une variable définie comme :

$$D_s(A_j, B_j) = \frac{l_a + l_b - 2 \cdot \text{inters}}{l_s} \tag{6.7}$$

où

inters = longueur$|A_j \cap B_j|$ est la longueur de l'intersection entre A_j et B_j.

REMARQUES

Il faut faire attention au problème d'indétermination posé par les deux cas suivants :
- $A_j = B_j$,
- les valeurs simples (qui sont des intervalles dégénérés).

Variables ordinale et nominale

Pour les variables ordinale et nominale Y_j, la distance $D(A_j, B_j)$ est définie en terme de composantes D_s et D_c seulement, ce qui revient à dire que

$$D(A_j, B_j) = D_s(A_j, B_j) + D_c(A_j, B_j) \tag{6.10}$$

et dans ce cas :

◇ la **composante due à l'étendue** est

$$D_s(A_j, B_j) = \frac{|l_a - l_b|}{l_s} \tag{6.11}$$

où

- l_a = le nombre de catégories de $A_j = |A_j|$,
- l_b = le nombre de catégories de $B_j = |B_j|$,
- inters = le nombre de catégories de $A_j \cap B_j = |A_j \cap B_j|$,
- l_s = le nombre de catégories de $A_j \cup B_j = |A_j \cup B_j| = l_a + l_b -$ inters,

et

◇ la **composante due au contenu** est :

$$D_c(A_j, B_j) = \frac{l_a + l_b - 2 \cdot \text{ inters}}{l_s} \tag{6.12}$$

A nouveau, la mesure proposée satisfait les propriétés d'une dissimilarité déterminée.

6.2.2 L'approche d'Ichino et Yaguchi (U-2, U-3, U-4)

En 1994, Ichino et Yaguchi [3] ont proposé une autre mesure de dissimilarité entre deux objets symboliques du type

$$D(a, b) = \sum_{j=1}^{p} D(A_j, B_j).$$

Cette distance sera appelée U-2.

Ils ont d'abord défini deux opérateurs cartésiens, $\oplus$ et $\otimes$ qui sont appliqués au couple de sous-ensemble (A_j, B_j) ; ils sont définis ci-dessous :

L'addition cartésienne

◇ Pour des variables réelles, entières et ordinales de type intervalle $A_j = [a_{j\ell}, a_{ju}]$ et $B_j = [b_{j\ell}, b_{ju}]$, l'**addition cartésienne** est définie par

$$A_j \oplus B_j = [\min(a_{j\ell}, b_{j\ell}), \max(a_{ju}, b_{ju})] \tag{6.13}$$

◇ Pour des variables nominales, l'**addition cartésienne** devient l'union de A_j et B_j, ie

$$A_j \oplus B_j = A_j \cup B_j \tag{6.14}$$

La multipliation cartésienne

La **multipliation cartésienne** de deux sous-ensembles $A_j, B_j \subseteq \mathcal{Y}_j$ est définie comme suit

$$A_j \otimes B_j = A_j \cap B_j \tag{6.15}$$

et est valable pour tous les types de variables.

La mesure de dissimilarité résultante

A partir de ces opérations cartésiennes, il est maintenant possible de développer la mesure de dissimilarité d'Ichino et Yaguchi, notée $\phi(A_j, B_j)$, pour une paire de sous-ensembles A_j, B_j appartenant au j$^{\text{ème}}$ espace d'observation $\mathcal{Y}_j$:

$$\phi(A_j, B_j) = |A_j \oplus B_j| - |A_j \otimes B_j| + \gamma(2 \cdot |A_j \otimes B_j| - |A_j| - |B_j|) \tag{6.16}$$

où

- $0 \leq \gamma \leq 0.5$ est un paramètre prédéfini.
- $|A_j|$ dénote soit la longueur de l'intervalle A_j (si la j$^{\text{ème}}$ variable est continue) soit le nombre d'éléments possibles inclus dans l'ensemble A_j.

Le paramètre $\gamma \in [0, 0.5]$ joue un rôle important dans cette définition. En fait, quand A_j et B_j sont des intervalles, γ contrôle l'effet de "proximitè intérieure" (inner-side nearness) et de "proximitè extérieure" (outer-side nearness) entre A_j et B_j sur la distance.

□ Si γ atteint sa valeur minimale, c'est-à-dire 0, la fonction de dissimilarité devient :

$$\phi(A_j, B_j) = |A_j \oplus B_j| - |A_j \otimes B_j| \tag{6.17}$$

et si A_j et B_j sont des intervalles disjoints, alors $\phi(A_j, B_j)$ reflète seulement la proximité extérieure et donc la distance entre les deux intervalles.

□ Si $\gamma = 0.5$, la fonction de dissimilarité $\phi(A_j, B_j)$ devient :

$$\phi(A_j, B_j) = |A_j \oplus B_j| - (|A_j| + |B_j|)/2. \tag{6.18}$$

Le choix de la valeur de γ est laissée à l'utilisateur.

La mesure de dissimilarité ϕ est une métrique équivalente.

Définition 6.2.1 *Une dissimilarité d est dite* **paire** *si* $\forall a, b \in E$

$$d(a, b) = 0 \Rightarrow d(a, c) = d(b, c) \ \ \forall a, b \in E$$

Combinaison des dissimilarités associées à chaque composante

Soient donnés deux objets a et b définis comme dans (6.1) et (6.2). Il est possible de définir la distance de Minkowski généralisée d'ordre q $(q \geq 1)$ comme :

$$d_q(a,b) = \left(\sum_{j=1}^{p} \phi(A_j, B_j)^q \right)^{1/q} \tag{6.19}$$

où toutes les variables Y_j peuvent être exprimées avec différentes unités de mesures.

L'inconvénient principal de cette formulation est la dépendance de la distance aux unités de mesure choisies. Par exemple, la distance entre deux années consécutives peut être 1 si l'unité choisie est une année, alors qu'il peut être 365 si l'unité choisie est un jour. Dans le but de résoudre ce problème, il est suggéré de normaliser la dissimilarité ϕ entre deux composantes comme suit :

$$\psi(A_j, B_j) = \frac{\phi(A_j, B_j)}{|\mathcal{Y}_j|} \tag{6.20}$$

où $|\mathcal{Y}_j|$ est la longueur du domaine $\mathcal{Y}_j$ si la j$^{\text{ème}}$ variable est continue, ou le nombre de valeurs possibles du domaine $\mathcal{Y}_j$ si la j$^{\text{ème}}$ variable est entière ou catégorique. En normalisant de la sorte, $0 \leq \psi(A_j, B_j) \leq 1$.

L'application de cette normalisation à la mesure de dissimilarité entre deux variables produit une nouvelle formulation de la distance entre deux objets symboliques booléens qui sera appelée $U-3$ et est définie par :

$$d_q(a,b) = \left(\sum_{j=1}^{p} \psi(A_j, B_j)^q \right)^{1/q}. \tag{6.21}$$

Une extension supplémentaire de la fonction distance définie ci-dessus peut être obtenue en introduisant des poids $w_j > 0$ vérifiant la relation $\sum_{j=1}^{p} w_j = 1$. L'introduction de poids permet de contrôler l'importance relative des variables introduites dans la fonction.

La distance de Minkowski modifiée correspondante devient :

$$d_q(a,b) = \left(\sum_{j=1}^{p} w_j [\psi(A_j, B_j)]^q \right)^{1/q} \tag{6.22}$$

et est appelée $U-4$. Cette mesure de dissimilarité vérifie les propriétés d'une métrique et les relations $0 \leq d_q(a,b) \leq 1$.

Remarquons que la distance euclidienne pondérée généralisée est obtenue en posant q=2 dans la définition (6.22).

6.2.3 Mesures de dissimilarité de De Carvalho

En 1994, 1996 et 1998, De Carvalho [3] propose deux extensions de la mesure de dissimilarité d'Ichino et Yaguchi.
La première extension concerne les fonctions ψ et ϕ qui furent créées par Ichino et Yaguchi avec la métrique de Minkowski généralisée. De Carvalho combine plusieurs fonctions, appelées fonctions de comparaison (FC), avec une fonction d'agrégat (FA), comme la distance de Minkowski.
La seconde extension introduit deux types de dépendances logiques entre différentes variables Y_j qui permettent de différencier les objets symboliques booléens.

Les fonctions de comparaison

Les fonctions de comparaison peuvent être introduites comme suit :
Supposons

$$a = [Y_1 \in A_1] \wedge [Y_2 \in A_2] \wedge \ldots \wedge [Y_p \in A_p]$$

et

$$b = [Y_1 \in B_1] \wedge [Y_2 \in B_2] \wedge \ldots \wedge [Y_p \in B_p],$$

deux objets symboliques sans contraintes (c'est-à-dire deux objets symboliques booléens avec aucune dépendance logique entre les variables).

De Carvalho suggère le calcul de la fonction de comparaison de chaque variable Y_j sur base des indices d'accord et de désaccord résumés dans le tableau suivant :

	accord	désaccord	total
accord	$\alpha = \mu(A_j \cap B_j)$	$\beta = \mu(A_j \cap C(B_j))$	$\mu(A_j)$
désaccord	$\chi = \mu(C(A_j) \cap B_j)$	$\delta = \mu(C(A_j) \cap C(B_j))$	$\mu(C(A_j))$
total	$\mu(B_j)$	$\mu(C(B_j))$	$\mu(O_j)$

où pour chaque sous-ensembles $V_j \subseteq \mathcal{Y}_j$

$$\mu(V_j) = \begin{cases} |V_j| & \text{si } \mathcal{Y}_j \text{ est une variable entière, nominale ou ordinale} \\ |\overline{v_j} - \underline{v_j}| & \text{si } \mathcal{Y}_j \text{ est continu et } V_j = [\underline{v_j}, \overline{v_j}] \text{ est un intervalle,} \end{cases}$$

et

$$C(V_j) = \mathcal{Y}_j - V_j$$

est l'ensemble complémentaire de V_j sur le domaine $\mathcal{Y}_j$.

Après avoir défini ces indices d'accord et de désaccord, De Carvalho a défini les fonctions de comparaison suivantes comme une extension des mesures de similarité définies pour les variables binaires classiques (avec valeurs 0 et 1).

S_i	fonction de comparaison	Rang	Propriété	=0 pour	=1 pour
S_1	$\frac{\alpha}{\alpha+\beta+\chi}$	$[0,1]$	métrique	$A_k \cap B_k = \varnothing$	$A_k = B_k$
S_2	$\frac{2\alpha}{2\alpha+\beta+\chi}$	$[0,1]$	semi-métrique	$A_k \cap B_k = \varnothing$	$A_k = B_k$
S_3	$\frac{\alpha}{\alpha+2(\beta+\chi)}$	$[0,1]$	métrique	$A_k \cap B_k = \varnothing$	$A_k = B_k$
S_4	$\frac{1}{2}[\frac{\alpha}{\alpha+\beta}+\frac{\alpha}{\alpha+\chi}]$	$[0,1]$	semi-métrique	$A_k \cap B_k = \varnothing$	$A_k = B_k$
S_5	$\frac{\alpha}{\sqrt{(\alpha+\beta)(\alpha+\chi)}}$	$[0,1]$	semi-métrique	$A_k \cap B_k = \varnothing$	$A_k = B_k$

Chaque fonction de similarité $S_i (i = 1, \ldots,)$ génère une fonction de dissimilarité correspondante d_i, e.g.

$$d_i = 1 - S_i. \tag{6.23}$$

La fonction de similarité S_i (ou de dissimailarité d_i) satisfait les propriétés suivantes :

□ Pour une variable réelle Y_j, la fonction de similarité S_i et la fonction de dissimilarité d_i sont invariantes pour tout changement d'échelle qui résulte d'une transformation linéaire de Y_j.

□ Les fonctions de similarité S_1, S_2, S_3 sont équivalentes, ce qui est encore vrai pour les fonctions de dissimilarité d_1, d_2, d_3.

Agrégat des fonctions de comparaison d_i

Comme Ichino et Yaguchi l'ont proposé pour les distances ϕ et ψ, De Carvalho sélectionne une des dissimilarités d_i pour chaque variable Y_j et les combine avec une fonction d'agrégat f comme, par exemple, la métrique de Minkowski. Il en résulte la dissimilarité globale :

$$d_q^i(a,b) = \sqrt[q]{\sum_{j=1}^{p} [w_j \, d_i(A_j, B_j)]^q} \tag{6.24}$$

avec $i \in \{1, \ldots, 5\}$. Cette distance est appelée SO-1.

Nous pouvons prouver ce qui suit :

Soient d_a et d'_a obtenus en applicant la même fonction d'agrégat f à deux mesures de dissimilarités, d_i et d'_i.
Si, pour chaque variable Y_j, d_i est équivalent à d'_i alors d_a est équivalent à d'_a.

Utilisation d'une mesure de comparaison normalisée

En 1996, De Carvalho [3] a aussi proposé une autre fonction de comparaison ψ' en combinaison avec une fonction f d'agrégat appropriée. Cette fonction de comparaison est définie comme une normalisation supplémentaire de la fonction ϕ d'Ichino et Yaguchi :

$$\psi'(A_j, B_j) = \frac{\phi(A_j, B_j)}{\mu(A_j \oplus B_j)}. \tag{6.25}$$

Cette fonction satisfait les propriétés suivantes :

□ $0 \leq \psi'(A_j, B_j) \leq 1$;

□ Pour une variable Y_j de valeur réelle, ψ' n'est pas affectée par un changement d'échelle dû à une transformation linéaire de Y_j.

□ ψ' est une distance équivalente (voir définition 6.2.1.).

En procédant avec la même fonction d'agrégat f avec la métrique de Minkowski généralisée pondérée, De Carvalho obtient la distance suivante entre deux objets symboliques booléens :

$$d'_q(a, b) = \sqrt[q]{\sum_{j=1}^{p} [\frac{1}{p} \psi'(A_j, B_j)]^q} \tag{6.26}$$

Notons que dans ce cas, nous avons choisi des poids égaux pour toutes les variables Y_j, c'est-à-dire que le poids associé à chaque variable est $w_j = \frac{1}{p}$. Cette distance est appelée SO-2.

Une approche de description potentielle pour les objets symboliques sans contraintes

La seconde extension de la fonction de distance d'Ichino et Yaguchi proposée par De Carvalho suppose le calcul de la distance d'Ichino et Yaguchi sans utiliser une fonction d'agrégat. Cela a été accompli en utilisant le concept de "description potentielle" (DP), notée $\pi(a)$, qui est utilisée pour calculer la fonction de comparaison ϕ.

Pour un objet symbolique booléen (sans contraintes) $a = [Y_1 = A_1] \wedge [Y_2 \in A_2] \wedge \ldots \wedge [Y_p \in A_p]$, $\pi(a)$ est défini comme le volume du produit cartésien $A_1 \times A_2 \times \ldots \times A_p$. Plus précisément, $\pi(a)$ est défini comme :

$$\pi(a) = \prod_{j=1}^{p} \mu(A_j), \tag{6.27}$$

où

$$\mu(A_j) = \begin{cases} |A_j| & \text{si } \mathcal{Y}_j \text{ est une variable entière, nominale ou ordinale} \\ |\overline{a_j} - \underline{a_j}| & \text{si } \mathcal{Y}_j \text{ est continu et } A_j = [\underline{a_j}, \overline{a_j}] \text{ est un intervalle,} \end{cases}$$

Pour calculer l'extension de De Carvalho pour les objets symboliques booléens, il est nécessaire d'étendre le définition des opérateurs cartésiens expliqués dans la section 6.2.2.

de la façon suivante :

□ Addition $\oplus$

$$a \oplus b = [Y_1 \in A_1 \oplus B_1] \wedge [Y_2 \in A_2 \oplus B_2] \wedge \ldots \wedge [Y_p \in A_p \oplus B_p] \tag{6.28}$$

où $A_j \oplus B_j$ est la définition donnée par Ichino et Yaguchi.

□ Multiplication $\otimes$

$$a \otimes b = [Y_1 \in A_1 \otimes B_1] \wedge [Y_2 \in A_2 \otimes B_2] \wedge \ldots \wedge [Y_p \in A_p \otimes B_p] \tag{6.29}$$

où $A_j \otimes B_j$ est la définition donnée par Ichino et Yaguchi.

A partir de ces extensions, la distance d'Ichino et Yaguchi entre deux objets symboliques booléens devient :

$$d_1'(a,b) = \pi(a \oplus b) - \pi(a \otimes b) + \gamma(2 \cdot \pi(a \otimes b) - \pi(a) - \pi(b)) \tag{6.30}$$

où $0 \leq \gamma \leq 0.5$, comme auparavant. Cette distance est appelée SO-3.

Il est possible de normaliser $d_1'(a,b)$ de deux façons différentes, et obtenir les distances suivantes

$$d_2'(a,b) = \frac{\pi(a \oplus b) - \pi(a \otimes b) + \gamma(2 \cdot \pi(a \otimes b) - \pi(a) - \pi(b))}{\pi(a^E)} \tag{6.31}$$

où $a^E = [Y_1 \in \mathcal{Y}_1] \wedge [Y_2 \in \mathcal{Y}_2] \wedge \ldots \wedge [Y_p \in \mathcal{Y}_p]$ et,

$$d_3'(a,b) = \frac{\pi(a \oplus b) - \pi(a \otimes b) + \gamma(2 \cdot \pi(a \otimes b) - \pi(a) - \pi(b))}{\pi(a \oplus b)} \tag{6.32}$$

appelées respectivement SO-4 et SO-5.

Ces mesures de dissimilarités satisfont les propriétés suivantes

1. $d_1'(a,b)$ et $d_2'(a,b)$ sont des mesures de dissimilarité définies mais l'inégalité triangulaire n'est pas vérifiée. De plus, d_1' et d_2' sont équivalentes et $d_2'(a,b)$ est normalisée entre 0 et 1.
2. $d_3'(a,b)$ est une métrique et est normalisée entre 0 et 1.
3. Pour p variables $Y_j \in \mathbb{R}$, l'ordre quasi défini par $(a,b) \preceq_{d_1} (a',b') \Leftrightarrow d_1'(a,b) \leq d_1'(a',b')$ sur $\mathbb{R}$ n'est pas affecté par des transformations linéaires des Y_j.
4. d_2' et d_3' ne sont pas affectées par des transformations linéaires des variables continues.

6.2.4 Mesure de dissimilarité pour les BSO avec contraintes

De Carvalho a également considéré les objets symboliques avec contraintes, c'est-à-dire des objets dont les variables peuvent dépendre l'une de l'autre.

Il y a deux types de dépendance entre les variables :

1. *Dépendance hiérarchique* (mère-fille) : Cette dépendance considère le cas où une variable Y_i ne peut pas être appliquée si une autre variable Y_j prend des valeurs dans un sous-ensemble $S_j \subseteq \mathcal{Y}_j$. Cette sorte de dépendance est une dépendance logique et peut s'écrire sous forme de la règle :

$$r_1 : \textbf{Si } [Y_j \in S_j] \textbf{ alors } [Y_i = NA]. \tag{6.33}$$

2. *Dépendance logique* : Ce cas se produit, par exemple, si un sous-ensemble $S_j \in \mathcal{Y}_j$ de variable Y_j dépend d'un autre sous-ensemble $S_i \in \mathcal{Y}_i$ d'une autre variable et peut se mettre sous forme de règle comme suit :

$$r_2 : \textbf{Si } [Y_j \in S_j] \textbf{ alors } [Y_i \in S_i]. \tag{6.34}$$

Exemple 6.2.2 *Exemple de dépendance hiérarchique.*
L'objet symbolique suivant :

$$a = [Sexe \in \{Masculin,\ Féminin\}] \wedge [Enceinte \in \{Oui,\ Non\}]$$

est contraint par la dépendance hiérarchique entre les variables Sexe et Enceinte exprimée comme suit :

$$\textbf{Si } [Sexe = Masculin] \textbf{ alors } [Enceinte = NA].$$

Exemple 6.2.3 *Exemple de dépendance logique.*
L'objet symbolique suivant :

$$b = [Age \in [0, 90]] \wedge [Taille \in [0.40, 2.00]\}]$$

est contraint par la dépendance logique entre les variables Age et Taille exprimée par :

$$\textbf{Si } [Age \in [0, 10]] \textbf{ alors } [Taille \in [0.40, 1.60]].$$

Etant donné que nous n'utiliserons pas ce type d'objets, nous ne développerons pas la mesure introduite par De Carvalho. A titre d'indication, cette mesure s'appelle C-1.

6.3 Dissimilarités pour des distributions de probabilité

Supposons que nous ayons une matrice de données symboliques $\underline{X} = (x_{ij})_{nxp}$.
Nous allons fournir des méthodes pour comparer les éléments ξ_{kj} et $\xi_{\ell j}$ de la j$^{\text{ème}}$ colonne de $\underline{X}$ dans le cas où Y_j est une variable modale.

Rappelons que, pour une variable modale Y_j qui est définie sur un ensemble d'individus $E = \{x_1, \ldots, x_n\}$, les entrées $\xi_{kj} = Y_j(x_k)$ et $\xi_{\ell j} = Y_j(x_\ell)$ de la matrice de données symboliques $\underline{X}$ sont des mesures, distributions de probabilité ou poids sur l'espace d'observation $\mathcal{Y}_j$ de Y_j qui est typiquement :

(a) un espace euclidien $\mathcal{Y}_j = \mathbb{R}^s$ de dimension s, donnée,

(b) un ensemble fini $\mathcal{Y}_j = \{1, \ldots, n\}$ de catégories, ou

(c) un ensemble infini comptable $\mathcal{Y}_j = \{0, 1, 2, 3, \ldots\}$ d'états ou niveaux.

Dans la suite et afin d'éviter des indices multiples, nous écrirons $\mathcal{Y} = \mathcal{Y}_j$ et $P = x_{kj} = Y_j(x_k)$ et $Q = x_{\ell j} = Y_j(x_\ell)$ pour les mesures.

Nous supposons ici que P et Q sont les distributions de probabilités sur le domaine $\mathcal{Y}$ qui sont caractérisées par :

(a) dans le cas d'un espace euclidien $\mathcal{Y}_j = \mathbb{R}^t$:

des distributions $p(y)$ et $q(y)$, respectivement, définies pour $y \in \mathbb{R}^t$ (e.g., une densité de distribution normale) telle que, pour n'importe quel ensemble $B \subset \mathbb{R}^t$, nous avons $P[B] = \int_B p(y)dy$ et $Q[B] = \int_B q(y)dy$, et

(b),(c) dans le cas d'un espace discret (e.g. $\mathcal{Y} = \{0, 1, 2, \ldots\}$:

des fonctions de probabilité $p(y) = P[\{y\}] = P[\widetilde{Y} = y]$ et $q(y) = Q[\{y\}] = Q[\widetilde{Y} = y]$, où $\widetilde{Y}$ dénote une variable aléatoire discrète (classique) possible de distributions P et Q, respectivement.

Nous décrivons ci-dessous diverses mesures $d(P, Q)$ qui caractérisent les diverses distributions p et q. Ces mesures sont bien connues de la théorie de probabilité classique et de la statistique. En probabilité, elles décrivent, par exemple, une différence pondérée entre les densités p et q et en statistique, elles mesurent la difficulté dans la distinction entre les distributions P et Q quand on effectue un test d'hypothèse dans lequel P est l'hypothèse nulle et Q, l'hypothèse alternative.

Pour n'importe laquelle de ces mesures, la dissimilarité entre deux données symboliques $P = x_{kj}$ et $Q = x_{\ell j}$ est alors définie par :

$$d^j_{k\ell} = d(Y_j(x_k), Y_j(x_\ell)) = d(x_{kj}, x_{\ell j}) = d(P, Q). \tag{6.35}$$

6.3.1 Mesures de divergence : Le cas général

Les index les plus classiques pour mesurer la déviation entre deux distributions de probabilité P et Q de densité p et q (pour une variable aléatoire $\widetilde{Y}$) sont des cas spéciaux de la ϕ-divergence ou I-divergence proposée par Csiszár (1967) qui utilise la proportion (vraisemblance) $\lambda(y) = q(y)/p(y)$ des densités p et q :

$$d(P,Q) = d(P,Q,\phi) = E_p[\phi(\lambda(\widetilde{Y}))]$$
$$= \int_{\mathcal{Y}} \phi(\lambda(y))dP(y) = \int_{\mathcal{Y}} \phi(\frac{q(y)}{p(y)})p(y)d\mu(y). \qquad (6.36)$$

où

- $\phi(.)$ est une fonction convexe à valeurs réelles définie sur $\mathbb{R}^+$ telle que $\phi(0) = 1$
- μ est soit la mesure de Lebesgue (cas (a)) soit la mesure de comptage (cas (b) et (c)) sur $\mathcal{Y}$.

Bien que cet index soit non-symétrique pour P et Q (voir dans la section destinée au cas spéciaux), l'inégalité de Jensen est satisfaite, c'est-à-dire que :

$$d(P,Q) = E_p[\phi(\lambda(\widetilde{Y}))] \geq \phi(E_p[\lambda(\widetilde{Y})])$$
$$= \phi\left(\int_{\mathcal{Y}} \frac{q(y)}{p(y)} \cdot p(y)d\mu(y)\right) = \phi(1) = 0 \qquad (6.37)$$

et donc

$$d(P,q) \geq 0$$

ce que doit vérifier un coefficient de dissimilarité. La relation $d(P,P) = 0$ pour $Q = P$ est également vérifiée.

Dans ce qui suit, nous listons quelques indices de dissimilarité majeurs résultant d'un choix spécial de la fonction $\phi(.)$. Les cas suivants seront traités pour tous les choix de ϕ :

(1) deux distributions continues P et Q ;

(2) deux distributions discrètes P et Q ;

(3) deux distributions normales de dimension t , $P = \mathcal{N}_t(\mu_k, \Sigma_k)$ et $Q = \mathcal{N}_t(\mu_\ell, \Sigma_\ell)$ de moyennes $\mu_k, \mu_\ell \in \mathbb{R}^t$ et de matrices de covariance définies positives Σ_k, Σ_ℓ ;

(4) deux distributions uni-dimensionnelles sur $\mathbb{R}^1$ constantes sur les intervalles $I_1, \ldots, I_m$ et $J_1, \ldots, J_M$ de $\mathbb{R}^1$ respectivement telle que

$$p(y) = \sum_{s=1}^{m} p_s.1_{I_s}(y) \text{ avec } y \in \mathbb{R}^1 \qquad (6.38)$$

$$q(y) = \sum_{t=1}^{M} q_t.1_{J_t}(y) \text{ avec } y \in \mathbb{R}^1 \qquad (6.39)$$

avce des poids $p_s, q_t > 0$, respectivement. Si nous notons par $K_{st} = I_s \cap J_t$ l'intersection des deux intervalles de p et q (plusieurs d'entre eux seront vides en pratique)

et par $\lambda_1(K_{st})$ la longueur (mesure de Lebesgue) de l'intervalle K_{st}, alors d(P,Q) est donnée par :

$$d(P,Q) = \sum_{s=1}^{m}\sum_{t=1}^{M}\int_{K_{st}} \log\left\{\frac{q_s}{p_s}\right\}\cdot p_s dy$$

$$= \sum_{s=1}^{m}\sum_{t=1}^{M} \log\left\{\frac{q_s}{p_s}\right\}\cdot p_s \cdot \lambda_1(K_{st}) \tag{6.40}$$

6.3.2 Mesures de divergence : Cas spéciaux

Il existe un grand nombre de choix pour la fonction convexe $\phi(\lambda)$ qui mènent à des index de divergence et de dissimilarité différents. Dans ce qui suit, nous allons étudier quelques cas spéciaux.

Information de discrimination de Kullback-Leibler

Cet index est obtenu pour un choix de $\phi(\lambda) = \lambda \cdot \log\lambda$ avec $\lambda > 0$ et $\phi(0) = 0$. Dans le cas continu, nous obtenons

$$d(P,Q) = I(P,Q) = \int_{\mathbb{R}^t} \log\left\{\frac{q(y)}{p(y)}\right\}\cdot q(y)dy \tag{6.41}$$

et est aussi appelé le I-coefficient ou divergence de Kullback-Leibler ou distance relative ou informative.

Similairement, dans le cas discret, nous avons

$$d(P,Q) = I(P,Q) = \sum_{y\in\mathcal{Y}} \log\left\{\frac{q(y)}{p(y)}\right\}\cdot q(y) \tag{6.42}$$

et pour deux distributions normales t-dimensionnelles, nous trouvons

$$d(P,Q) = \frac{1}{2}\left(\|\mu_k - \mu_\ell\|^2_{\Sigma_k^{-1}} + tr(\Sigma_k^{-1}\Sigma_\ell - I_t) + \log\left(\frac{|\Sigma_k|}{|\Sigma_\ell|}\right)\right) \tag{6.43}$$

où

$$\|\mu_k - \mu_\ell\|^{-1}_{\Sigma_k^{-1}} = (\mu_k - \mu_\ell)'\Sigma_k^{-1}(\mu_k - \mu_\ell) \tag{6.44}$$

est la distance de Mahalanobis.
Pour des matrices de covariance identiques, c'est-à-dire, $\Sigma_k = \Sigma_\ell = \Sigma$,

$$d(P,Q) = \frac{1}{2}\left(\|\mu_k - \mu_\ell\|^2_{\Sigma}\right) \tag{6.45}$$

et on retrouve le carré de la distance euclidienne, $d_{kl} = \frac{1}{2}\|\mu_k - \mu_\ell\|^2$ pour une matrice de covariance qui est l'identité, i.e. $\Sigma = I_t$. Remarquons que si $\Sigma_k = \Sigma_\ell = \Sigma$, le coefficient de dissimilarités est symétrique ce qui n'est pas le cas sinon.

Le J-coefficient ; ϕ-divergence symétrisée

Le coefficient de Kullback-Leibler peut être symétrisé. Il est alors appelé J-coefficient et s'exprime de la manière suivante :

$$J(P,Q) = I(Q|P) + I(P|Q). \tag{6.46}$$

On peut facilement montrer que cette mesure est ϕ-divergente pour la fonction $\phi(\lambda) = (\lambda - 1) \cdot \log \lambda$.

Pour deux distributions normales, nous obtenons par (6.44) que

$$J(P,Q) = \frac{1}{2}\left(\|\mu_k - \mu_\ell\|^2_{\Sigma_k^{-1}} + \|\mu_k - \mu_\ell\|^2_{\Sigma_\ell^{-1}} + tr(\Sigma_k^{-1}\Sigma_\ell - I_t) + tr(\Sigma_\ell^{-1}\Sigma_k - I_t)\right) \tag{6.47}$$

et pour deux distributions de types histogrammes,

$$J(P,Q) = \sum_{s=1}^{m}\sum_{t=1}^{M} \log\left\{\frac{q_t}{p_s}\right\} \cdot (q_t - p_s) \cdot \lambda_1(K_{st}). \tag{6.48}$$

Plus généralement, si $d(P,Q) = d(P,Q,\phi)$ est une ϕ-divergence arbitraire de fonction convexe $\phi(\lambda)$, alors en interversant les arguments P et Q, nous obtenons une autre mesure de divergence $\widetilde{d}$:

$$\begin{aligned} d(Q,P;\phi) &= \int_y \phi\left(\frac{p(y)}{q(y)}\right) q(y) d\mu(y) \\ &= \int_y \widetilde{\phi}\left(\frac{q(y)}{p(y)}\right) p(y) d\mu(y) = d(P,Q;\widetilde{\phi}) =: \widetilde{d}(P,Q) \end{aligned} \tag{6.49}$$

où la fonction conjuguée $\widetilde{\phi}(\lambda) = \lambda.\phi(1/\lambda)$.

De plus, l'index symétrisé est

$$\begin{aligned} d^*(P,Q) &= d(P,Q;\phi) + d(Q,P;\phi) \\ &= d(P,Q;\phi) + d(P,Q;\widetilde{\phi}) = d(P,Q;\phi^*) \end{aligned} \tag{6.50}$$

qui est a nouveau une ϕ-divergence de fonction convexe $\phi^*(\lambda) = \phi(\lambda) + \widetilde{\phi}(\lambda)$.

Donc, pour tous les coefficients de ϕ-divergence non symétriques d, il existe une version symétrique d^* qui satisfait tous les axiomes d'une mesure de dissimilarité, mais pas nécessairement l'inégalité triangulaire.

La divergence χ^2

Cette divergence utilise la fonction $\phi(\lambda) = (\lambda-1)^2$ qui induit une mesure de dissimilarité qui compare la différence au carré $|p(y) - q(y)|^2$ et la densité $p(y)$. On obtient donc dans le cas continu

$$d(P,Q) = \int_{\mathbb{R}^t} \left(\frac{q(y)}{p(y)} - 1\right)^2 p(y)dy = \int_{\mathbb{R}^t} \frac{|p(y) - q(y)|^2}{p(y)} dy \tag{6.51}$$

et dans le cas discret

$$d(P,Q) = \sum_{y \in \mathcal{Y}} \frac{|p(y) - q(y)|^2}{p(y)}. \tag{6.52}$$

Ces mesures de dissimilarité sont appelées χ^2-divergence.
Pour deux distributions normales, nous trouvons

$$d(P,Q) = \frac{|\Sigma_k \, \Sigma_\ell^{-1}|}{|2 \cdot \Sigma_k \Sigma_\ell^{-1} - I_t|^{1/2}} \cdot \exp\left\{\frac{1}{2}L\right\}^{-1} \tag{6.53}$$

avec $L = \left(\|2 \cdot \Sigma_\ell^{-1}\mu_\ell - \Sigma_k^{-1}\mu_k\|^2_{(2\cdot\Sigma_\ell^{-1} - \Sigma_k^{-1})^{-1}} + \|\mu_k\|^2_{\Sigma_k^{-1}} - 2 \cdot \|\mu_\ell\|^2_{\Sigma_\ell^{-1}}\right)$.

Notons que la matrice $2 \cdot \Sigma_\ell^{-1} - \Sigma_k^{-1}$ doit être définie positive. Pour des matrices de covariance identiques, $\Sigma_\ell = \Sigma_k = \Sigma$, la formule est réduite à

$$d(P,Q) = \exp\left\{\|\mu_k - \mu_\ell\|^2_{\Sigma^{-1}}\right\} - 1 \tag{6.54}$$

et pour deux histogrammes, nous obtenons

$$d(P,Q) = \sum_{s=1}^{m}\sum_{t=1}^{M} \frac{|p_s - q_t|^2}{p_s} \cdot \lambda_1(K_{st}). \tag{6.55}$$

Le coefficient d'Hellinger

Diverses mesures de dissimilarités ont été définies en utilisant le coefficient d'Hellinger qui résulte d'un choix spécial de la fonction convexe $\phi(\lambda) = \phi_s(\lambda) = \lambda^s$, où s est un exposant positif tel que $0 < s < 1$. Ce coefficient est donné par

$$\begin{aligned} d^{(s)}(P,Q) &= \int_{\mathbb{R}^t} \left(\frac{q(y)}{p(y)}\right)^s p(y)dy \\ &= \int_{\mathbb{R}^t} q(y)^s \cdot p(y)^{1-s} dy. \end{aligned} \tag{6.56}$$

Remarquons que cette fois, la fonction ϕ_s est concave. Mais la transformation triviale $\widetilde{\phi}_s = 1 - \phi_s$ donne une mesure de divergence équivalente et possède une fonction $\widetilde{\phi}_s$ convexe. Le cas discret est donné par

$$d^{(s)}(P,Q) = \sum_{y \in \mathcal{Y}} q^s(y) \cdot p^{(1-s)}(y). \tag{6.57}$$

Le coefficient Hellinger est un *coefficient de similarité* vérifiant les propriétés suivantes :

1. $0 \leq d^{(s)}(P,Q) \leq 1$,
2. $d^{(s)}(P,Q) = 1$ pour $P = Q$ et
3. $d^{(s)}(P,Q) = 0$ si P et Q sont des supports disjoints.

Le cas spécial d'Hellinger pour $s = 1/2$ mène au coefficient symétrique suivant

$$d^{(s)}(P,Q) = \int_{\mathbb{R}^t} \sqrt{p(y)q(y)}dy \tag{6.58}$$

et est appelé distance de Bhattacharya par Fukunaga (1972).

Pour deux normales, nous avons

$$d^{(s)}(P,Q) = |sI_t + (1-s)\Sigma_k^{-1}\Sigma_\ell|^{-s/2} \cdot |(1-s)I_t + s\Sigma_\ell^{-1}\Sigma_k|^{-(1-s)/2} \cdot \exp\left\{\frac{1}{2}M\right\}^{-1} \tag{6.59}$$

$$\text{avec } M = \left(\|s\Sigma_\ell^{-1}\mu_\ell + (1-s)\Sigma_k^{-1}\mu_k\|^2_{(s\Sigma_\ell^{-1}+(1-s)\Sigma_k^{-1})^{-1}} - s\|\mu_\ell\|^2_{\Sigma_\ell^{-1}} - (1-s)\|\mu_k\|^2_{\Sigma_k^{-1}}\right)$$

Si les matrices de covariance sont identiques, c'est-à-dire si $\Sigma_\ell = \Sigma_k = \Sigma$, l'équation 6.59 est réduit à la distance de Mahalanobis transformée, c'est-à-dire :

$$d^{(s)}(P,Q) = \exp\left\{-s(1-s) \cdot \|\mu_k - \mu_\ell\|^2_{\Sigma^{-1}}\right\} \tag{6.60}$$

Pour cette mesure de similarité, plusieurs coefficients de dissimilarité ont été dérivés. Ceux-ci jouent un rôle important en statistique. On trouve par exemple, la distance de Chernoff d'ordre s qui est utilisée dans le module DISS et qui s'exprime comme suit :

$$d(P,Q) = -\log d^{(s)}(P,Q) \tag{6.61}$$

ou l'information d'ordre s

$$d(P,Q) = \frac{\log d^{(s)}(P,Q)}{s-1} \tag{6.62}$$

proposée par Rényi et également présente dans le module DISS.

Distance de Variation et la distance L_2

Finalement pour un choix de $\phi(\lambda) = |\lambda - 1|$, nous obtenons un coefficient de dissimilarité symétrique appelé distance de variation et qui s'exprime

$$d_1(P,Q) \int_{\mathbb{R}^t} |p(y) - q(y)|dy. \tag{6.63}$$

Dans le cas discret, nous obtenons donc

$$d_1(P,Q)\sum_{y\in\mathcal{Y}}|p(y)-q(y)|. \tag{6.64}$$

Il s'agit de la distance L_1 de Minkowski pour les densités $p(y)$ et $q(y)$.
Similairement, nous pouvons utiliser la distance L_2 de Minkowski donnée par

$$d_2(P,Q)\int_{\mathbb{R}^t}|p(y)-q(y)|^2dy. \tag{6.65}$$

Dans le cas de deux distributions normales dans $\mathbb{R}^t$, nous trouvons

$$d_2(P,Q)=\frac{1}{2^t\pi^{t/2}}\left(\frac{1}{|\Sigma_k|^{1/2}}+\frac{1}{|\Sigma_\ell|^{1/2}}\right)-\frac{2}{(2\pi)^{t/2}|\Sigma_k+\Sigma_\ell|^{1/2}}\cdot\exp\left\{\frac{1}{2}N\right\} \tag{6.66}$$

avec $N=\left(\|\Sigma_k^{-1}\mu_k+\Sigma_\ell^{-1}\mu_\ell\|^2_{(\Sigma_k^{-1}+\Sigma_\ell^{-1})}-\|\mu_k\|^2_{\Sigma_k^{-1}}-\|\mu_\ell\|^2_{\Sigma_\ell^{-1}}\right)$
qui, pour des matrices de covariance identiques, est réduit à

$$d_2(P,Q)=\frac{1}{2^t\pi^{t/2}|\Sigma|^{1/2}}\cdot\left(1-\exp\{\frac{1}{4}\|\mu_k-\mu_\ell\|^2_{\Sigma^{-1}}\}\right) \tag{6.67}$$

Nous pouvons à nouveau généraliser et donc obtenir une distance L_p qui s'exprime dans le cas discret comme

$$d_p(P,Q)=\sum_{y\in\mathcal{Y}}|p(y)-q(y)|^p. \tag{6.68}$$

6.3.3 Mesures de dissimilarité entre deux PSO

Soit deux objets symboliques

$$a=[Y_1\in A_1]\wedge[Y_2\in A_2]\wedge\ldots\wedge[Y_p\in A_p]$$

et

$$b=[Y_1\in B_1]\wedge[Y_2\in B_2]\wedge\ldots\wedge[Y_p\in B_p]$$

où chaque variable Y_j est une variable modale et prend ses valeurs sur un domaine $\mathcal{Y}_j$ et A_j, B_j sont des sous-ensembles de $\mathcal{Y}_j$. Une fonction de dissimilarité entre a et b peut être construite en agrégeant les coefficients de dissimilarité entre les distributions de probabilité en généralisant et pondérant la métrique de Minkowski comme suit :

$$d_p(a,b)=\sqrt[p]{\sum_{k=1}^{p}[c_k d^*(A_k,B_k)]^p} \tag{6.69}$$

où $\forall k \in \{1, \ldots, p\}, c_k > 0$ est le poids associé à chaque variable. Les poids vérifient la relation $\sum_{k=1}^{p} c_k = 1$.
En posant $P = A_k$ et $Q = B_k$, $m(P, Q)$ est un coefficient de dissimilarité entre les distributions de probabilité qui est un de ceux vus dans la section précedente, c'est-à-dire :

- J-coefficient, noté J :

$$J(P, Q) = I(Q|P) + I(P|Q)$$
$$\text{où } I(Q|P) = \sum_{y \in \mathcal{Y}} \log \left\{ \frac{q(y)}{p(y)} \right\} \cdot p(y)$$

- χ^2-divergence symétrisée, noté $CHI2$, et s'exprimant :

$$d^*(P, Q) = \left(\sum_{y \in \mathcal{Y}} \frac{|p(y) - q(y)|^2}{p(y)} + \sum_{y \in \mathcal{Y}} \frac{|q(y) - p(y)|^2}{q(y)} \right)$$

- la distance de Chernoff symétrisée, notée $CHER$:

$$d^*(P, Q) = -\log d^{(s)}(P, Q) - \log d^{(s)}(Q, P)$$
$$\text{où } 0 < s < 1 \text{ et } d^{(s)}(P, Q) = \sum_{y \in \mathcal{Y}} q^s(y) \cdot p^{1-s}(y).$$

- la distance de Rènyi symétrisée, notée REN :

$$d^*(P, Q) = \frac{\log d^{(s)}(P, Q)}{s - 1} + \frac{\log d^{(s)}(Q, P)}{s - 1}$$

- la distance L_p de Minkowski, notée LP :

$$d^*(P, Q) = \sum_{y \in \mathcal{Y}} |p(y) - q(y)|^p$$
$$\text{avec } p > 0.$$

La distance (6.69) est appelée PU-1 et est présente dans le module DISS.

Alternativement, les coefficients de dissimilarité peuvent être agrégé à travers un produit. Par conséquent, en prenant toutes les précautions nécessaires et en considérant la mesure L_p de Minkowski, nous obtenons la mesure de dissimilarité normalisée suivante entre des PSOs :

$$d'_p(a, b) = 1 - \frac{\prod_{i=1}^{p} \left(\sqrt[p]{2} - \sqrt[p]{\sum_{y_i} |p(y_i) - q(y_i)|^p} \right)}{(\sqrt[p]{2})^n} = 1 - \frac{\prod_{i=1}^{p} (\sqrt[p]{2} - \sqrt[p]{L_p})}{(\sqrt[p]{2})^n} \qquad (6.70)$$

où chaque y_i correspond à la valeur du domaine de la $i^{\text{ème}}$ variable.
Remarquons que cette mesure de dissimilarité est symétrique et normalisée sur $[0, 1]$.

Evidemment, $d'_p(a, b) = 0$ si $a = b$ et $d'_p(a, b) = 1$ si les deux objets sont complètement différents.

6.4 Le module DISS

Dans les sections précédentes, nous venons d'expliquer comment mesurer la dissimilarité entre deux composantes d'un objet symbolique et entre des objets symboliques booléens ou probabilistiques. Nous allons à présent présenter le module qui nous permet d'élaborer la matrice de distance entre de tels objets. Ce module est très important dans notre démarche car il va nous permettre d'obtenir en sortie la matrice de distance avec laquelle nous travaillerons dans la suite. Nous verrons dans le chapitre 7, comment nous sommes parvenu à modifier le programme SCLUST dans le but de mettre en entrée cette matrice et d'obtenir une classification ainsi que les indices de Milligan et Cooper.

6.4.1 Introduction

Le module DISS, créé par l'université de Bari, a été développé dans le cadre du projet ASSO et est très facilement applicable à partir du logiciel SODAS. Cette méthode a été élaborée pour comparer des objets symboliques dans le but de quantifier les corrélations existantes entre eux, de les classifier ou de les discriminer. Nous utilisons ce module afin de classifier des objets symboliques décrits par des variables de tout type et des combinaisons entre ces variables.

6.4.2 L'entrée du module DISS

Comme nous l'avons déjà signalé, le module DISS évalue le degré de dissimilarité entre les objets d'un ensemble de données symboliques. L'utilisateur doit d'abord insérer la base de données, qui est soit un fichier .SDS ou un fichier.XML et ensuite, il optera pour la méthode DISS. On peut voir une illustration de la page du logiciel SODAS sur la figure 6.1.

Nous savons aussi qu'un objet symbolique est généralement décrit par une collection de variables valuées et modales. L'utilisateur doit choisir les variables avec lesquelles il veut travailler pour calculer les dissimilarités entre les objets symboliques.

Selon les variables choisies, les mesures de dissimilarité que nous avons étudiées dans les sections 6.2 et 6.3 peuvent être sélectionnées. Si nous sommes dans le cas mixte où il y a des variables modales ou valuées, l'utilisateur doit choisir une mesure de dissimilarité pour les BSO et une pour les PSO. La matrice de dissimilarité finale est obtenue en combinant les deux matrices calculées en utilisant les deux mesures sélectionnées.

Nous savons aussi qu'un objet symbolique est généralement décrit par une collection de variables valuées et modales. L'utilisateur doit choisir les variables avec lesquelles il veut travailler pour calculer les dissimilarités entre les objets symboliques.

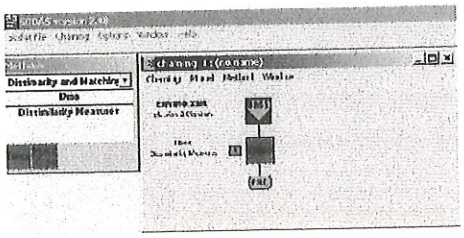


FIG. 6.1 – L'entrée du module DISS

Selon les variables choisies, les mesures de dissimilarité que nous avons étudiées dans les sections 6.2 et 6.3 peuvent être sélectionnées. Si nous sommes dans le cas mixte où il y a des variables modales ou valuées, l'utilisateur doit choisir une mesure de dissimilarité pour les BSO et une pour les PSO. La matrice de dissimilarité finale est obtenue en combinant les deux matrices calculées en utilisant les deux mesures sélectionnées.

Les mesures de dissimilarité peuvent être calculées en associant le même poids ($\frac{1}{p}$ si le nombre de variables est p) à chaque variable. Mais l'utilisateur peut également décider d'associer un poids différent à chaque variable sélectionnée.

L'utilisateur doit également spécifier le fichier de sortie (NOM + CHEMIN) qui est soit un fichier de type .SDS ou .XML. En ce qui nous concerne, nous avons favorisé le type .XML. Nous verrons pourquoi nous avons fait ce choix dans le chapitre 7.
Ces choix sont automatiquement enregistrés dans un fichier paramètre (fichier.pad) qui assure la communication entre SODAS et le module DISS. Ce fichier paramètre est composé de six sections.

La première section contient :

- Le fichier SODAS (CHEMIN + NOM) qui est l'entrée de la chaîne
  (i.e. SDS-IN = "C :\ SODAS \ DISS \ enviro.SDS")
- Le fichier SODAS (CHEMIN + NOM) qui contient la sortie de DISS
  (i.e. SDS-OUT = "C :\ SODAS \ DISS \ output1.XML")
- Le fichier log (CHEMIN + NOM)qui contient le lod du calcul de DISS
  (i.e. LOG = "C :\ SODAS \ DISS \ exemple1.LOG")
- Le fichier de sortie (CHEMIN + NOM) qui possède la matrice de dissimilarité
  (i.e. OUT = "C :\ SODAS \ DISS \ exemple1.LST")

La seconde section contient les poids éventuellement associés à chaque variable sélectionnée.
Remarque : Si une liste de poids est introduite, la somme des poids doit valoir 1.

La troisième section, aussi appelée section BSO, contient tous les paramètres qui per-

mettent de calculer la mesure de dissimilarité sélectionnée pour les BSOs.

Similairement à la troisième section, la quatrième appelée aussi section PSO, contient les paramètres concernant la mesure de dissimilarité sélectionnée pour les PSOs.

La cinquième section contient la liste des variables sélectionnées

Finallement, la sixième section contient la modalité ($\sum$ ou $\prod$) qui permet de combiner la matrice de dissimilarité obtenues dans le cas où des variables modales ou non modales ont été sélectionées dans la section 5.

Exemple 6.4.1 *Exemple complet d'un fichier paramètre :*

SDS-IN = "C :\Program Files\DECISIA\SODAS version 2.0\bases\merovingianmulti.sds"
SDS-OUT = "C :\Program Files\DECISIA\SODAS version 2.0\bases\merovingienmultiU2.xml"
LOG = "C :\Program Files\DECISIA\SODAS version 2.0\filieres\CG5P1Y01.LOG"
OUT = "C :\Program Files\DECISIA\SODAS version 2.0\filieres\CG5P1Y01.LST"

PROC-DISTANCE

: :—Weights for selected variables
WSEL = EQUIV
: :—List of parameters for BSO
SIMFUN = U-2
POWER = 2
GAMMA = 0.50
: :—List of parameters for PSO
SIMFUN = PU-1
COMPFUN = J
POWER = 2
: :—Selected variables
SELECT = 1-6
: :—Combine option
COMBINE = SUMM

◇

La sortie du module DISS

La sortie de DISS est une matrice triangulaire. Cela est dû à la propriété de symétrie que possède les mesures de dissimilarité. En effet, pour deux objets symboliques a et b, nous avons la relation suivante

$$d(a,b) = d(b,a).$$

Ce fichier de sortie est écrit dans un nouveau fichier SODAS. Il s'agit d'une copie du fichier d'entrée auquel s'ajoute la matrice de dissimilarité. Il peut être trouvé à l'adresse donnée dans SDS-OUT, c'est-à-dire si nous prenons l'exemple 6.4.1, à l'adresse :

"C :\Program Files\DECISIA\SODAS version 2.0\bases\merovingienmultiU2.xml"

D'autres sorties sont également disponibles, notamment un fichier texte contenant uniquement la matrice de dissimilarité.

La méthode DISS

Les mesures de dissimilarité qui permettent d'évaluer le degré de dissimilarité entre chaque paire d'objets symbolique sont détaillées ci-après. Elles ont toutes été introduites dans les sections 6.2 et 6.3.

◇ Les mesures de dissimilarité pour les BSO

Soit deux objets symboliques booléens a et b :

$$a = [Y_1 \in A_1] \wedge [Y_2 \in A_2] \wedge \ldots \wedge [Y_p \in A_p]$$

$$b = [Y_1 \in B_1] \wedge [Y_2 \in B_2] \wedge \ldots \wedge [Y_p \in B_p]$$

définis par p variables.
Chaque variable Y_j prend ses valeurs dans un domaine $\mathcal{Y}_j$ et A_j et B_j sont des sous-ensembles de $\mathcal{Y}_j$. Nous avons vu qu'il était possible de définir une mesure de disimilarité entre deux objets booléens a et b en agrégeant les valeurs de dissimilarité calculées indépendamment pour chaque variable simple (dissimilarité composante par composante). Une fonction d'agrégat classique est la métrique de Minkowski généralisée. Cependant d'autres classes de mesures sont définies pour des BSO. Celles-ci sont basées sur la notion de description potentielle, $\pi(a)$, défini par $A_1 \times A_2 \times \ldots \times A_P$. Pour cette classe de mesures, la décomposition composante par composante n'est pas nécessaire, et donc aucune fonction d'agrégat n'est requise.

Les mesures de dissimilarité définies actuellement dans le module DISS sont :
- U-1 : la mesure de dissimilarité de Gowda et Diday,
- U-2 : la première formulation de la mesure de dissimilarité d'Ichino et Yaguchi,
- U-3 : la mesure de dissimilarité d'Ichino et Yaguchi normalisée,
- U-4 : la mesure de dissimilarité d'Ichino et Yaguchi normalisée et pondérée,
- SO-1 : la mesure de dissimilarité de De Carvalho,
- SO-2 : la mesure de dissimilarité d'Ichino et Yaguchi étendue par De Carvalho,
- SO-3 : la première mesure de dissimilarité basée sur la description potentielle,
- SO-4 : la seconde mesure de dissimilarité basée sur la description potentielle,
- SO-5 : la mesure de dissimilarité basée sur la description potentielle normalisée,

– C-1 : la mesure de dissimilarité de De Carvalho pour les BSOs avec contraintes.

Elles sont reprises dans le tableau ci-après.

Nom	Mesure de dissimilarité par composantes	Mesure de dissimilarité globale
$U-1$	$D^{(j)}(A_j,B_j)=D_\pi(A_j,B_j)+D_s(A_j,B_j)$ $+D_c(A_j,B_j)$où D_π est dû à la position , D_s est dû à l'étendue et D_c est dû au contenu.	$d(a,b)=\sum_{i=1}^{p} D^{(j)}(A_j,B_j)$
$U-2$	$\phi(A_j,B_j)=\lvert A_j\oplus B_j\rvert-\lvert A_j\times B_j\rvert$ $+\gamma(2\cdot\lvert A_j\otimes B_j\rvert-\lvert A_j\rvert-\lvert B_j\rvert)$où$\otimes$ et $\oplus$ sont deux opérateurs cartésiens.	$d_q(a,b)=\sqrt[q]{\sum_{j=1}^{p}[\phi(A_j,B_j)]^p}$
$U-3$	$\psi(A_j,B_j)=\frac{\phi(A_j,B_j)}{\lvert\mathcal{Y}\rvert}$	$d_q(a,b)=\sqrt[q]{\sum_{j=1}^{p}[\psi(A_j,B_j)]^p}$
$U-4$	$\psi(A_j,B_j)=\frac{\phi(A_j,B_j)}{\lvert\mathcal{Y}\rvert}$	$d_q(a,b)=\sqrt[q]{\sum_{j=1}^{p} w_j[\psi(A_j,B_j)]^p}$
$SO-1$	$d_i(A_j,B_j)i=1,\ldots,5$	$d_q^i(a,b)=\sqrt[q]{\sum_{j=1}^{p}[w_j d_i(A_j,B_j)]^p}$
$SO-2$	$\psi'(A_j,B_j)=\frac{\phi(A_j,B_j)}{\mu(A_j\oplus B_j)}$	$d'_q(a,b)=\sqrt[q]{\sum_{j=1}^{p}\frac{1}{p}[\psi'(A_j,B_j)]^p}$
$SO-3$	néant	$d'_1(a,b)=\pi(A_j\oplus B_j)+\pi(A_j\otimes B_j)$ $-\gamma(2\pi(A_j\otimes B_j)-\pi(a)-\pi(b))$
$SO-4$	néant	$d'_2(a,b)=\frac{d'_1(a,b)}{\pi(a^E)}$ où $\pi(a^E)=[Y_1\in\mathcal{Y}_1]\wedge\ldots\wedge[Y_p\in\mathcal{Y}_p]$
$SO-5$	néant	$d'_3(a,b)=\frac{d'_1(a,b)}{\pi(a\oplus b)}$
$C-1$	non vu	

TAB. 6.1 – Mesures de dissimilarité pour les BSOsS

Nous allons maintenant revoir les différentes mesures utilisées dans DISS ainsi que leurs paramètres respectifs dans le tableau 6.2.

Mesure de dissimilarité	Paramètres	Contraintes	Valeurs par défaut
$U-1$	néant		
$U-2$	gamma ordre de la puissance	$[0, 0.5]$ $1 \ldots 10$	0.5 2
$U-3$	gamma ordre de la puissance	$[0, 0.5]$ $1 \ldots 10$	0.5 2
$U-4$	gamma ordre de la puissance poids par variables	$[0, 0.5]$ $1 \ldots 10$ Somme(poids) $= 1$	0.5 2 poids égaux
$SO-1$	fonction de comparaisons ordre de la puissance poids par variables	D_1, D_2, D_3, D_4, D_5 $1 \ldots 10$ Somme(poids) $= 1$	D_1 2 poids égaux
$SO-2$	gamma ordre de la puissance	$[0, 0.5]$ $1 \ldots 10$	0.5 2
$SO-3$	gamma	$[0, 0.5]$	0.5
$SO-4$	gamma	$[0, 0.5]$	0.5
$SO-5$	gamma	$[0, 0.5]$	0.5

TAB. 6.2 – Mesures de dissimilarités pour les BSOs ainsi que leurs paramètres

◇ Les mesures de dissimilarité pour les PSOs

Soit deux objets symboliques booléens a et b :

$$a = [Y_1 \in A_1] \wedge [Y_2 \in A_2] \wedge \ldots \wedge [Y_p \in A_p]$$

$$b = [Y_1 \in B_1] \wedge [Y_2 \in B_2] \wedge \ldots \wedge [Y_p \in B_p]$$

définis par p variables modales qui prennent des valeurs dans un domaine $\mathcal{Y}_j$ et A_j, B_j sont des sous-ensembles de $\mathcal{Y}_j$.
Une fonction de dissimilarité entre a et b peut être construite en agrégeant les coefficients de dissimilarité entre les distributions de probabilité en utilisant la métrique de Minkowski généralisée et pondérée (6.69). On retrouve cette mesure sous le nom de PU-1 dans le tableau 6.3 et elle s'exprime comme suit :

$$d_p(a, b) = \sqrt[p]{\sum_{k=1}^{p} [c_k d^*(A_k, B_k)]^p}.$$

Pour rappel, on utilise les fonctions de comparaison suivantes pour calculer le coefficient de dissimilarité entre des distributions de probabilité :

- J : J-coefficient,
- CHI2 : χ^2-divergence,
- CHER : la distance de Chernoff,
- REN : la distance de Rènyi et
- LP : la distance L_p de Minkowski

Il est également possible d'agréger les coefficients de dissimilarité en effectuant un produit. Par conséquent, on obtient la mesure PU-2 qui est :

$$d'_p(a,b) = 1 - \frac{\prod_{i=1}^{p}(\sqrt[p]{2} - \sqrt[p]{L_p})}{(\sqrt[p]{2})^n}$$

Nom	Mesure de dissimilarité par composantes		Mesure globale
PU-1	J-Coefficient	$J(P,Q) = I(Q\|P) + I(P\|Q)$ où $I(Q\|P) = \sum_{y\in\mathcal{Y}} \log\{\frac{q(y)}{p(y)}\} \cdot q(y)$	$d_q(a,b) = \sqrt[q]{\sum_{j=1}^{p}[c_k d^*(A_j,B_j)]^q}$
	χ^2-divergence symétrisée	$d^*(P,Q) = (\sum_{y\in\mathcal{Y}} \frac{\|p(y)-q(y)\|^2}{p(y)} + \sum_{y\in\mathcal{Y}} \frac{\|q(y)-p(y)\|^2}{q(y)})$	
	Distance de Chernoff symétrisée	$d^*(P,Q) = -\log d^{(s)}(P,Q) - \log d^{(s)}(Q,P)$ où $0 < s < 1$ et $d^{(s)}(P,Q) = \sum_{y\in\mathcal{Y}} q^s(y) \cdot p^{1-s}(y)$.	
	Distance de Rényi symétrisée	$d^*(P,Q) = \frac{\log d^{(s)}(P,Q)}{s-1} + \frac{\log d^{(s)}(Q,P)}{s-1}$	
	Distance L-p de Minkowski	$d^*(P,Q) = \sum_{y\in\mathcal{Y}} \|p(y) - q(y)\|^p$ où $p > 0$	
Pu-2	la distance L_p de Minkowski	$LP = \sum_{y\in\mathcal{Y}} \|p(y) - q(y)\|^p$	$d(a,b) = 1 - \frac{\prod_{i=1}^{p}(\sqrt[p]{2} - \sqrt[p]{L_p})}{(\sqrt[p]{2})^n}$

TAB. 6.3 – Mesures de dissimilarités disponibles pour les PSOs

Dans le tableau 6.4, nous présentons les paramètres présents dans chaque mesure et les valeurs possibles de ceux-ci ainsi que celles attribuées par défaut.

Mesure de dissimilarité	Paramètres	Contraintes	Valeurs par défaut
$PU-1$	fonction de comparaison ordre de la puissance s p poids par variables	$J, CHI2, CHER, REN, LP$ $1 \ldots 10$ $[0,1]$ $1 \ldots 10$ Somme(poids) $= 1$	J 2 0.5 1 poids égaux
En particulier			
$PU-1(J)$	ordre de la puissance poids par variables	$1 \ldots 10$ Somme(poids) $= 1$	2 poids égaux
$PU-1(CHI2)$	ordre de la puissance poids par variables	$1 \ldots 10$ Somme(poids) $= 1$	2 poids égaux
$PU-1(REN)$	s ordre de la puissance poids par variables	$[0,1]$ $1 \ldots 10$ Somme(poids) $= 1$	0.5 2 poids égaux
$PU-1(CHER)$	s ordre de la puissance poids par variables	$[0,1]$ $1 \ldots 10$ Somme(poids) $= 1$	0.5 2 poids égaux
$PU-1(LP)$	p ordre de la puissance poids par variables	$1 \ldots 10$ $1 \ldots 10$ Somme(poids) $= 1$	1 2 poids égaux
$PU-2$	p	$1 \ldots 10$	1

TAB. 6.4 – Mesures de dissimilarité pour les PSOs ainsi que leurs paramètres

Le module DISS dans le logiciel SODAS

Nous venons de voir les entrées et sorties du module DISS ainsi que les mesures de dissimilarité que nous pouvons y trouver. Nous allons maintenant expliquer brièvement comment utiliser ce logiciel.

L'utilisateur du logiciel choisit premièrement "Dissimilarity and matching" dans l'ensemble des méthodes disponibles. Il ouvre une chaîne existante ou il peut décider d'en créer une.
Nous allons par exemple prendre le cas où l'utilisateur décide de créer une nouvelle chaîne appelée Disstest associée au fichier SODAS enviro.xml. Ensuite, pour insèrer DISS dans la chaîne, il clique sur le bloc "BASE" et en choisissant "insert method". Un nouveau bloc vide s'ajoute dans la chaîne. L'utilisateur de SODAS sélectionne alors la méthode DISS et la porte sur le bloc vide.

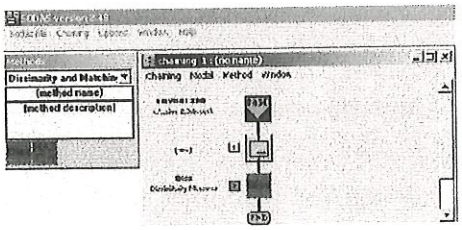


FIG. 6.2 – Première démarche du module DISS

En cliquant après sur le bloc contenant la méthode sélectionnée, il suffit de choisir le champ "parameter" et une nouvelle fenêtre apparaît. Cette fenêtre est donnée par la figure 6.3.

Enviro.xml contient 14 objets symboliques décrits par 17 variables mixtes, à la fois booléennes et modales. Dans la section "variables" de la nouvelle fenêtre, l'utilisateur choisit soit toutes les variables soit une partie de celles-ci.

L'utilisateur doit également choisir les mesures de dissimilarité avec lesquelles il souhaite travailler. Suivant les variables sélectionnées, il choisira une mesure pour les BSOs et/ou les PSOs dans la section "variables". Une illustration de la fenêtre est donnée à la figure 6.4

Remarque : Il ne faut pas oublier de sauver le rapport en .sds ou en .xml.
Finalement, l'utilisateur exécute la méthode DISS en choisissant "run method" dans le menu "method". Un nouveau bloc correspondant au fichier de rapport généré par DISS est ajouté à la chaîne. Celui-ci peut alors être analysé. Il est également possible de retrouver ce même rapport en format .sds ou .xml à l'endroit où il a été enregistré dans "save output file".
Remarquons que ces fichiers .sds et .xml peuvent être utilisé dans différents programmes. C'est ce que nous ferons en modifiant le programme SCLUST.

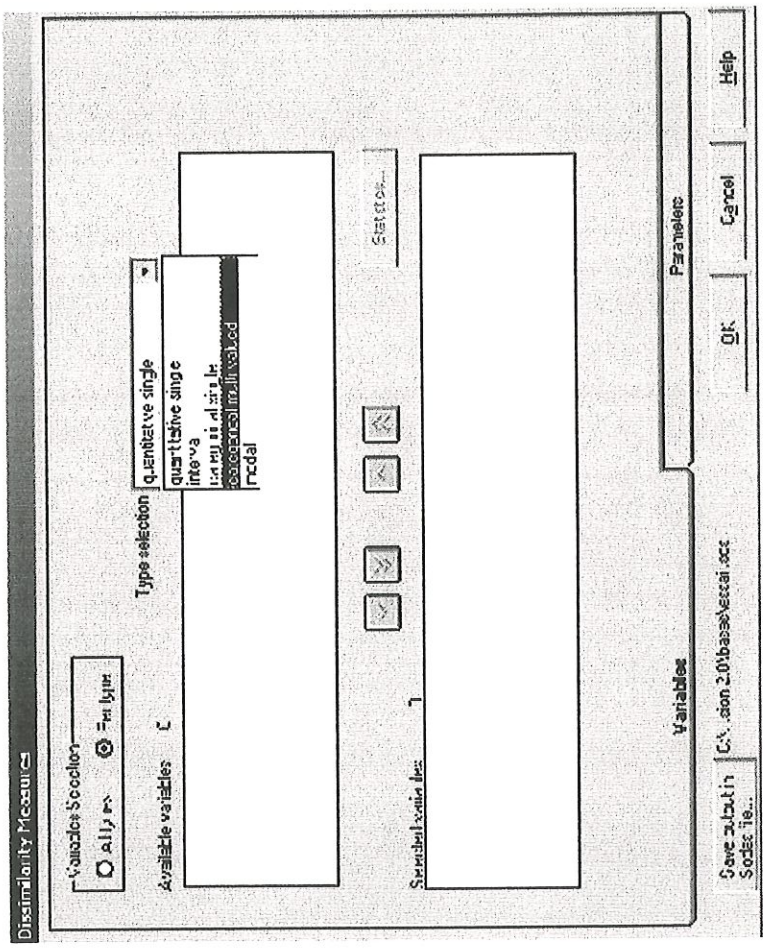


FIG. 6.3 – La fenêtre "parameter"

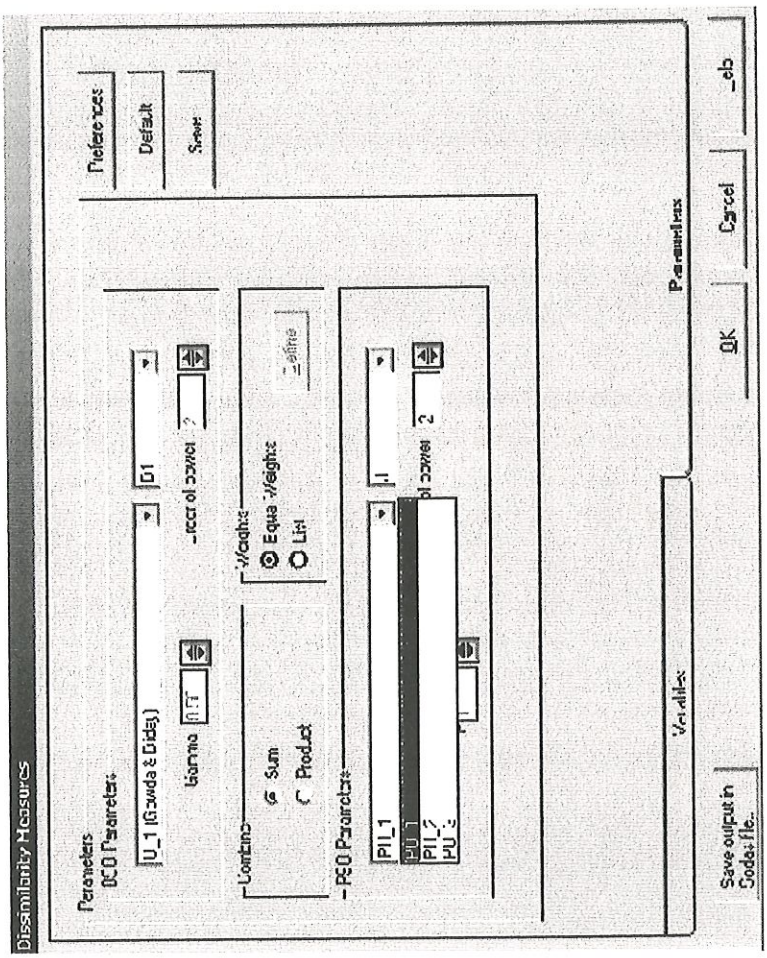


FIG. 6.4 – La fenêtre "variables"

Chapitre 7

Le programme SCLUST

7.1 Introduction

SCLUST est une méthode symbolique de classification en cours de développement réalisée dans le cadre du contrat européen ASSO. Le code C++ de ce programme a été réalisé par Yves Lechevallier de l'INRIA Rocquencourt.

SCLUST est basé sur la méthode des nuées dynamiques et détermine de façon itérative une série de partitions qui améliore à chaque étape un critère de classification basé sur des prototypes qui représentent les classes et une fonction de proximité pour affecter les objets symboliques aux classes.

SCLUST traite les trois différents types de variables symboliques, à savoir :
- les variables intervalles ;
- les variables multivaluées ;
- les variables modales.

ainsi que les variables quantitatives et qualitatives classiques.

Jusqu'ici il était possible d'obtenir une détermination du nombre de classes pour un seul type de variables à la fois. L'objectif de ce mémoire étant d'obtenir une détermination du nombre de classes pour toute combinaison de variables, nous allons rappeler brièvement comment fonctionne SCLUST et expliquer comment nous avons modifié ce programme afin d'obtenir une classification à partir d'une matrice de distance obtenue à l'aide du module DISS.

7.2 Brève description de SCLUST

7.2.1 Le fichier sodas

Le fichier sodas est généré à l'aide du module DB2SO du logiciel SODAS à partir d'une base de données présentée en format Excel ou Access. Il possède tout ce que SCLUST à besoin de connaître sur le jeu de données pour pouvoir effectuer la classification.
Celui-ci est divisé en plusieurs blocs. Ceux-ci sont minimum au nombre de six à savoir
CONTAINS qui liste tous les blocs du fichier SODAS.
FILE informe sur l'origine du fichier.
HEADER contient le nom du jeu de données, le nombre total d'individus et de variables dans le fichier, le nombre de variables de chaque type, ainsi que le nombre de données manquantes.
INDIVIDUALS donne à chaque individu un numéro et un libellé.
VARIABLES donne un numéro, son type et un libellé à chaque variable.
RECTANGLE MATRIX est la matrice de données symboliques.

Mais il peut y avoir des blocs supplémentaires tels que RULES , HIERARCHIE, ...

Dans les travaux réalisés précédemment ce type de fichier était le seul type en entrée. C'est ce type de fichier que nous avons mis en entrée dans le module DISS. Quand nous utiliserons SCLUST, nous prendrons un fichier .XML en entrée étant donné que cela n'a pas encore été réalisé.

Un fichier .XML contient exactement les mêmes informations qu'un fichier sodas, mais les fichiers .XML sont beaucoup plus lisibles.
Voici un exemple de fichier .XML.

CONTAINS

Files	Header	Individuals	Variables	Rules	Matrix	Dissimilarity Matrix	Matching Matrix	Dependence
YES	YES	YES	YES	NO	YES	NO	NO	NO

FILE DESCRIPTION
Creation Procedure : sds2xml(v2.1)
Create Date : Mon Feb 02 12 :02 :04 2004

HEADER

Title	auto
Sub Title	auto
Number of Individuals	21
Number of Variables	7
Number of interval variables	7
Number of categorical multi-valued variables	0
Number of not applicable (NA)	0
Number of dependences (Rules)	0
Number of taxonomies	0
Number of missing values (NULL)	0

LIST OF INDIVIDUALS

NUMERO	Short Name	Name
0	AA00	Alfa 145
1	AA01	Aston Martin
2	AA02	Audi A3
3	AA03	Audi A8
4	AA04	Bmw serie 3
5	AA05	Bmw serie 7
6	AA06	Ferrari
7	AA07	Punto
8	AA08	Focus
9	AA09	Honda NSK
10	AA10	Lamborghini
11	AA11	Lancia K
12	AA12	Maserati GT
13	AA13	Mercedes SL
14	AA14	Mercedes Classe E
15	AA15	Nissan Micra
16	AA16	Vectra
17	AA17	Porsche
18	AA18	Twingo
19	AA19	Rover 25
20	AA20	Skoda Fabia

LIST OF VARIABLES

Position	Short Name	Name	Type	NA	NULL	Description
1	AB00	Pression	intervalle	0	0	min: 16992, max: 460000
2	AC00	Cylindrée	intervalle	0	0	min: 973, max: 5992
3	AD00	Vitesse-Max	intervalle	0	0	min: 150, max: 335
4	AE00	Accélération	intervalle	0	0	min: 3.9, max: 17
5	AF00	Longueur	intervalle	0	0	min: 343, max: 516
6	AG00	Largueur	intervalle	0	0	min: 160, max: 204
7	AF00	Hauteur	intervalle	0	0	min: 111, max: 148

SYMBOLIC DATA TABLE

	Pression	Cylindrée	Vitesse-Max	Accélération	Longueur	Largeur	Hauteur
1	[27806,33596]	[1370,1910]	[185,211]	[8.3,11.2]	[406,406]	[171,171]	[143,143]
2	[260500,460000]	[5935,5935]	[298,306]	[4.7,5]	[465,467]	[183,192]	[124,132]
3	[40230,68838]	[1595,1781]	[189,238]	[6.8,10.9]	[415,415]	[174,174]	[142,142]
4	[123849,171417]	[2771,4172]	[232,250]	[5.4,10.1]	[503,503]	[188,188]	[144,144]
5	[45407,76392]	[1796,2979]	[201,247]	[6.6,10.9]	[447,447]	[174,174]	[142,142]
6	[104892,276792]	[2793,5397]	[228,240]	[7,8.6]	[498,512]	[186,186]	[143,143]
7	[240292,391692]	[3586,5474]	[295,298]	[4.5,5.2]	[476,476]	[192,192]	[130,130]
8	[19229,30885]	[1242,1910]	[155,170]	[12.2,14.3]	[380,384]	[166,166]	[148,148]
9	[27492,34092]	[1596,1753]	[185,193]	[10.8,11]	[415,415]	[170,170]	[143,143]
10	[205242,215242]	[2977,3179]	[260,270]	[5.7,6.5]	[414,414]	[175,175]	[129,129]
11	[413000,423000]	[5992,5992]	[335,335]	[3.9,3.9]	[447,447]	[204,204]	[111,111]
12	[58806,81306]	[1998,2959]	[212,220]	[8.9,9.2]	[469,469]	[183,183]	[146,146]
13	[155000,159500]	[3217,3217]	[280,290]	[5.1,5.7]	[451,451]	[182,182]	[131,131]
14	[132800,262500]	[2799,5987]	[232,250]	[6.1,9.7]	[447,447]	[181,181]	[129,129]
15	[69243,389405]	[1998,5439]	[222,250]	[5.7,9.7]	[482,482]	[180,180]	[144,144]
16	[18492,24192]	[998,1348]	[150,164]	[12.5,15.5]	[375,375]	[160,160]	[144,144]
17	[36492,49092]	[1598,2171]	[193,207]	[10.5,12.5]	[450,450]	[171,171]	[143,143]
18	[147704,246412]	[3387,3600]	[280,305]	[4.2,5.2]	[443,444]	[177,183]	[130,131]
19	[16992,23492]	[1149,1149]	[151,168]	[11.7,13.4]	[343,343]	[163,163]	[142,142]
20	[21492,33042]	[1119,1994]	[160,185]	[10.7,15]	[399,399]	[169,169]	[142,142]
21	[19519,32686]	[1397,1896]	[157,183]	[11.5,16.5]	[396,396]	[165,165]	[145,145]

7.2.2 Le fichier sclust.h

Dans le langage C++, une structure permet de regrouper plusieurs données de type différents, qui qualifient un même objet.

Dans le fichier Sclust.h, une structure appelée SCLUST est définie. Elle permet de stocker des données de différents types, et notamment :

- le nombre d'objets dans le fichier sodas ;
- le nombre de variables dans le fichier sodas ;
- le nombre d'objets sélectionnés ;
- le nombre de classes ;
- le nombre d'itérations ;
- le code de la distance ;
- le code de l'initialisation ;
- ⋮

Les valeurs prises par chacune des variables présentes dans la structure dépendent bien évidemment du fichier sodas et elles restent inchangées tout au long de l'exécution du programme.

7.2.3 Le fichier calcul_scluster.cpp

Ce fichier est le plus important de SCLUST dans le sens où il contient l'algorithme des nuées dynamiques dans le cas symbolique.

La fonction Init_scluster

Cette fonction est destinée à l'initialisation de la partition.

SCLUST permet en fait deux types d'initialisation :

1. une initialisation par une partition aléatoire et
2. une initialisation par prototypes.

◇ Initialisation par une partition aléatoire

La fonction Init_scluster, connaissant par la structure SCLUST le nombre d'individus sélectionnés (nb_objects) et le nombre de classes demandées (nb_class), elle associe aléatoirement à chaque individu un nombre réel compris entre 0 et (nb_class). Ce nombre est ensuite tronqué à l'entier supérieur, et c'est ce nombre qui détermine la classe initiale à laquelle appartient l'individu.

Ensuite, le nombre d'individus présents dans chaque classe ainsi que le prototype de chaque classe sont déterminés.

C'est la fonction Prot_scluster qui construit les différents prototypes de chaque classe à partir de ses individus. Les prototypes sont ensuite stockés dans une matrice Proto[][], où Proto[k][l] est le prototype de la modalité l relatif à la classe k.

La valeur du critère est ensuite calculée par la fonction W_scluster.

◇ **Initialisation par prototypes**

Si par contre, l'utilisateur choisit une initialisation par prototypes, la fonction associe aléatoirement à chaque classe un nombre réel compris entre 0 et nb_objects. Ce nombre est par la suite tronqué à l'entier supérieur qui détermine l'individu ou le prototype qui représentera la classe considérée.

Remarque : Deux classes distinctes ne peuvent être représentées par le même individu.

La fonction Affect_scluster

Chaque classe étant maintenant représentée par un prototype, cette fonction permet de trouver pour chaque individu le prototype qui lui est le plus proche au sens de la distance choisie par l'utilisateur, et affecte l'individu considéré à la classe représentée par ce prototype.

Cette fonction correspond à la fonction d'affectation f de la méthode des nuées dynamiques.

La fonction W_scluster

Dans la méthode des nuées dynamiques, vue au chapitre 4, il est nécessaire de définir un critère d'arrêt qui permet de mesurer l'adéquation entre toute partition et toute représentation de cette partition.

Dans SCLUST, le critère choisi est le suivant :

$$W(P,L) = \sum_{\ell=1}^{k} \sum_{x_i \in C_\ell} d^2(x_i, L_\ell)$$

où :
- k est le nombre de classes ;
- C_ℓ est la classe ℓ ;
- L_ℓ est le représentant de la classe ℓ.

Le critère cherché donc à minimiser la somme des écarts entre les individus de chaque classe et leur représentant, c'est-à-dire l'inertie intra-classes W du nuage de points.

La fonction Proto_scluster

Après avoir affecté chaque individu à la classe de laquelle il est le plus proche, il faut reconstruire le prototype de chaque classe, c'est-à-dire déterminer, sur base des individus présents dans chaque classe, le représentant de celle-ci.

La fonction Proto_scluster se charge de la construction de ces prototypes et elle correspond à la fonction de représentation g de la méthode des nuées dynamiques.

7.2.4 La fonction main_scluster

Cette fonction contient l'algorithme principal des nuées dynamiques. Nous allons voir comment sont implémentées les fonctions présentées précédemment.

Supposons que la structure SCLUST est initialisée. Voici les principales étapes de l'algorithme.

```
for (n = 1; n <= n_run; n + +) // boucle sur le nombre de runs
{

  Init_scluster(lis,para,M,vsel,classe, Nt, Proto,Pw);// initialisation de la méthode
  iter=0; n_diff=imax;
  while( (iter < iter_max) && (n_diff ! = 0)) // boucle sur les iterations
  {
    iter = iter + 1;

    /****** Etape d'affectation ******/

    n_diff=Affec_scluster(lis,para,M,vsel,classe,Proto,Pw);// Calcul de la partition
    ww=W_scluster(lis,para,M,vsel,classe,Pw,Proto); // Calcul du critère

    /****** Etape de représentation ******/

    Proto_scluster(lis,para,M,vsel,classe,Proto); // Calcul des prototypes
    ww=W_scluster(lis,para,M,vsel,classe,Pw,Proto); // Calcul du critère
  }; // FIN boucle sur les iterations

  /****** stockage des informations sur ce run******/

  Wopt[n]=ww;N_iter[n]=iter;Nb_classe[n]=0;
  for (k=1;k<=kmax;k++) Nt[k]=0;
  for (i=1;i<=imax;i++) Nt[classe[i]]++;
  for (k=1;k<=kmax;k++) if(Nt[k] ! = 0) Nb_classe[n]++;
```

```
/****** Si solution optimale stockage du run ******/

(...)
 if(ww < w_max)
  {
   max_run=n ;w_max=ww ;
   for (i = 1; i <= imax; i + +) classe_opt[i]=classe[i] ;
   for (k = 1; k <= kmax; k + +)
       {
          for (m = 0; m <= mod_max; m + +)
              {
                 Proto_opt[k][m]=Proto[k][m] ;
              } ;
       } ;
  } ;
} ; // FIN boucle sur le nombre de runs
```

Fixons un essai (run) de recherche de la partition optimale en k classes, noté n.

Premièrement, l'algorithme calcule la partition initiale et le prototype de chacune des classes par le biais de la fonction Init_scluster.

Ensuite, une boucle sur les itérations est effectuée. A chacune de ces itérations les fonctions Affect_scluster, Proto_scluster et W_scluster sont utilisées itérativement. Les itérations sont effectuées jusqu'à ce que soit le nombre d'itérations maximal est atteint ou soit le nombre de changements de classes d'affectation de tous les individus (n_diff) renvoyé par la fonction Affect_scluster est nul.

Après des informations sur ce run sont stockées, à savoir :
- la valeur optimale du critère dans W_opt[n] ;
- le nombre d'itérations effectuées pour obtenir la solution dans N_iter[n] ;
- le nombre d'individus par classe dans Nt[classe[i]] ;
- le nombre de classes dans Nb_classe[n].

Enfin, si le run considéré n donne une valeur du critère inférieure à celle des $n-1$ runs considérés jusqu'alors, ce run n est stocké, ainsi que :
- la valeur du critère ww dans wmax ;
- la partition optimale dans classe_opt[i] ;
- les prototypes de chaque classe dans Proto_opt[k][m].

7.2.5 Le fichier listing

Le fichier listing est le fichier dans lequel sont répertoriés tous les résultats obtenus par SCLUST sur un jeu de données précisé.

Pour commencer, SCLUST affiche les valeurs prises par les principales variables de la structure SCLUST définies par l'utilisateur. Ces valeurs sont fixées pour toute l'exécution du programme. La valeur de l'inertie totale est également imprimée.

Dans un second temps, les informations suivantes nous sont données :
- le numéro de chaque variable sélectionnée ;
- le nom de chaque variable ;
- le type de chaque variable ;
- les valeurs :

$$100 \cdot \frac{T_j}{T} \text{ et } 100\frac{W_j}{W}$$

  c'est-à-dire les contributions de chaque variable à l'inertie totale et à l'inertie intra-classes. Ces deux valeurs sont égales au début de la classification car l'ensemble des individus forme encore une seule classe.
- le poids de chaque variable.

Ensuite, pour chaque itération de chaque essai, sont indiqués le nombre de changements de classes d'affectation de tous les individus et la valeur du critère. Remarquons que la valeur du critère est décroissante.

A la fin de tous les essais, SCLUST est en mesure de nous fournir le run pour lequel la solution optimale a été obtenue ainsi que la valeur de cette solution.

L'édition de la partition optimale nous informe également sur la composition des classes formant la partition.

Viens ensuite la description de la partition, pour laquelle les quantités suivantes sont calculées :
- la dispersion du nuage de point qui est égale à l'inertie totale du nuage de points.
- le critère qui est égal à l'inertie intra-classes.
- le pourcentage de dispersion qui vaut :

$$100 \cdot \frac{T - W}{T} = 100 \cdot \frac{B}{T}.$$

Viens alors la description des variables. Pour chacune d'entre elles, les indices suivants sont calculés :
- COR(j)=$100 \cdot \frac{B_j}{T_j}$ qui représente la part d'inertie de la variable j prise en compte par la partition.
- CTR(j) = $100 \cdot \frac{B_j}{B}$ qui représente la contribution relative de chaque variable à l'inertie inter-classes B.

Par la suite, SCLUST nous fournit une description de chaque classe. Nous connaissons alors :
- la valeur T ;

- la valeur B, qui est définit par :

$$B = 100 \cdot \frac{T - W^{(\ell)}}{B}$$

et qui représente la contribution relative de chaque classe à l'inertie inter-classes.
- la valeur W, qui est définit par :

$$W = 100 \cdot \frac{W^{(\ell)}}{W}$$

et qui représente la contribution relative de chaque classe à l'inertie intra-classes.

L'édition des prototypes variable par variable nous informe sur :
- le prototype de chaque classe.
- COR qui est ici calculé pour une classe ℓ, par

$$COR(j, \ell) = \frac{W_j^{(\ell)}}{T_j},$$

- CTR qui est ici calculé, pour une classe ℓ, par la formule

$$CTR(j, \ell) = \frac{W_j^{(\ell)}}{W_j}.$$

Finalement, nous disposons pour chaque essai :
- du nombre d'itérations effectuées ;
- du nombre de classes trouvées ;
- de la valeur optimale du critère.

7.3 Adaptation du programme SCLUST

L'objectif principal de ce mémoire est d'adapter les méthodes de détermination du nombre de classes de Milligan et Cooper au programme SCLUST pour n'importe quelle combinaison de variables symboliques, intervalles, multivaluées et modales.

Nous allons pour cela calculer la matrice de dissimilarités entre les différents objets à l'aide du module DISS. C'est le fichier contenant cette matrice que nous utiliserons dans le programme SCLUST. Auparavant, cette matrice était calculée dans le programme par le biais de la matrice de données. Nous avons choisi de sauver les matrices de distances du module DISS dans des fichiers .xml. En effet, tous les travaux réalisés les années précédentes mettaient en entrée un fichier sodas, nous avons voulu innover et mettre un type de fichiers plus facilement lisible.

7.3.1 Construction des hiérarchies de partitions

Comme nous l'avons vu au chapitre 3, les quatre méthodes de classification hiérarchiques nécessitent le calcul d'une matrice de distances ou de dissimilarités entre individus à partir de laquelle les critères agglomératifs associés à chaque méthode effectuent les regroupements.

Cette matrice de dissimilarités est directement calculée dans le module DISS. Il nous suffit simplement de la lire. Cela est réalisable grâce aux lignes de code suivantes :

```
tab_dist d;
d=new tab_dist(imax);
d=(tab_dist*)M → getdist();
for (i=1; i<=imax; i++)
    for (int j=1; j<=imax; j++)
    {
        d->getd(i,j);
        DT[i][j]=d->getd(i,j);
    };.
```

A partir de cette matrice de distance qui est stockée dans la variable DT, les hiérarchies de partitions liées à chaque méthode hiérarchique peuvent alors être construites suivant les différents critères agglomératifs.

7.3.2 Méthodes de détermination du nombre de classes de Milligan et Cooper

Les méthodes de détermination du nombre de classes de Milligan et Cooper ont été initialement programmées par A.D. Gordon pour être appliquées à quatre méthodes hiérarchiques. Programmé initiallement en Fortran 77, le code de ce programme a été "traduit" en C++ par S. Delogne. Il l'a ensuite inséré dans le programme SCLUST pour calculer les cinq "meilleures" méthodes de Milligan et Cooper (Calinski et Harabaz, Duda et Hart, C-index, Gamma et Beale) appliquées aux hiérarchies de partitions générées par les quatre méthodes hiérarchiques, et ce aux différents niveaux des hiérarchies créées.

De plus les méthodes de Calinski et Harabaz, du C-index et Gamma ont été adaptées pour qu'elles puissent être appliquées aux partitions successives générées par la méthode de classification non-hiérarchique SCLUST.

Et donc les différentes méthodes de détermination du nombre de classes sont testées sur cinq méthodes de classification.

Le programme que nous avons conçu se trouve sur cd à la fin de ce rapport et est nommé DISS.

Deuxième partie

Applications

Chapitre 8

Introduction

Le but de cette partie est d'appliquer les méthodes de détermination du nombre de classes aux données symboliques de tout type et à des combinaisons de différents types. Dans un premier temps, nous analyserons les résultats que nous avons obtenus sur les jeux de données exploités par S. Delogne et S. Collès et les comparerons ensuite avec ceux obtenus auparavant.

Notre méthode fût de calculer la matrice de distances à l'aide du programme DISS. Grâce à cette matrice, il était facilement possible de nous servir des méthodes classiques de détermination du nombre de classes étudiées au chapitre 5. Ces méthodes ont été testées tant sur les hiérarchies de partitions produites par les quatre méthodes hiérarchiques que sur les partitions produites par le programme SCLUST.

Nous allons premièrement rappeler les méthodes de classification et les méthodes de détermination du nombre de classes utilisées.

Les méthodes de classification sont :
- la méthode du saut minimum ;
- la méthode du saut maximum ;
- la méthode du centroïde ;
- la méthode de Ward et
- SCLUST.

Les méthodes de détermination du nombre de classes sont les suivantes :
- la méthode de Calinski et Harabasz (M1) ;
- la méthode de Duda et Hart (M2) ;
- la méthode du C-index (M3) ;
- la méthode Gamma (M4) et
- la méthode de Beale (M5).

Nous avons testé les méthodes présentées ci-dessus sur les jeux de données suivants :

1. Pour les variables de type intervalle :
   - des données artificielles avec trois classes hypersphériques ;
   - des données artificielles avec deux classes allongées ;
   - des données artificielles avec deux classes emboîtées ;
   - des données artificielles sans structure ;
   - des données basées sur les formes de Breiman et
   - des données réelles reprenant les températures dans soixante villes chinoises pendant les douzes mois de l'année.
2. Pour les variables modales :
   - des données regroupant sur trois années les ventes de 13 magasins de vêtements et accessoires, répartis dans six pays différents et
   - des données décrivant les habitudes de consommation des ménages au Royaume-Uni.
3. Pour les variables multivaluées :
   - des données issues des boucles mérovingiennes ;
   - des données traitant d'aniamus.
4. Pour des variables intervalles et multivaluées :
   - des données reprenant dix caractéristiques automobiles mesurées sur trente-trois voitures.

Chapitre 9

Présentation des résultats

Les résultats obtenus par les méthodes de détermination du nombre de classes de Milligan et Cooper sur les hiérarchies de partitions générées par les quatres méthodes de classification hiérarchiques sont présentées sous la forme d'un tableau à double entrée :

		M1	M2	M3	M4	M5
Saut minimum						
Saut maximum						
Centroïde						
Ward						

TAB. 9.1 – Tableau utilisé pour représenter les résultats des méthodes de Milligan et Cooper

Chaque ligne du tableau correspond à une des quatre méthodes de classification hiérarchiques et les cinq dernières colonnes se rapportent aux cinq méthodes de détermination du nombre de classes de Milligan et Cooper. Par exemple, un "3" à l'intersection de la ligne associée à la méthode du saut minimum et de la colonne M2 indique que la méthode de Duda et Hart appliquée à la hiérarchie de partitions générée par la méthode du saut minimum suggère l'existence de trois classes dans les données. Nous prendrons comme convention d'inscrire :

- $\times$ si les indices ne sont pas suffisamment significatifs et
- ? si les indices sont significatifs mais sont fort semblables l'un de l'autre.

La deuxième colonne du tableau indique si les méthodes de classification retrouvent les classes naturelles pour un k fixé et correspondant au bon nombre de classes. UN "+" témoigne d'une classification correspondant aux classes naturelles, un "-" au cas contraire.

Pour ce qui est des partitions générées par SCLUST, le tableau aura une forme simplifiée dans le sens où les méthodes de Duda et Hart (M2) et de Beale (M5) ne sont pas applicables aux méthodes de classification non hiérarchiques.

		M1	M3	M4
SCLUST				

TAB. 9.2 – Tableau utilisé pour représenter les résultats des méthodes de Milligan et Cooper

De plus, nous associerons à chaque jeu de données les valeurs des indices des cinq méthodes de détermination du nombre de classes de Milligan et Cooper pour une méthode de classification particulière et pour différents nombres k de classes. Le tableau est de la forme :

Méthode de classification	M1	M2	M3	M4	M5
k=8					
k=7					
k=6					
k=5					
k=4					
k=3					
k=2					
k=1					

TAB. 9.3 – Valeurs des indices de Milligan et Cooper pour la méthode de classification considérée

Par exemple, la valeur de Gamma (M4) lorsque les données sont partitionnées en trois classes sera indiquée à l'intersection de la ligne $k = 3$ et de la colonne M4.

Bien évidemment le tableau associé à la méthode SCLUST aura la même forme si ce n'est que les colonnes se rapportant aux méthodes de Duda et Hart (M2) et de Beale (M5) n'apparaîtront pas.

Remarque :

Nous n'avons pas utilisé la normalisation des variables dans nos programmes. En effet, si nous utilisions cette normalisation, les résultats obtenus étaient très mauvais.

Par exemple, pour le jeu de données avec trois classes hypersphériques, nous constations qu'aucune des méthodes de classification ne retrouvait la partition naturelle. De plus, les méthodes de Milligan et Cooper ne retrouvaient pas le bon nombre de classes dans les données.

Chapitre 10

Les variables de type intervalle

10.1 Introduction

Le but de ce chapitre était de comparer les résultats obtenus à l'aide du module DISS avec ceux obtenus par S. Delogne [1] sur les jeux de données étudiés par celui-ci. Au vu des très mauvais résultats que nous avons obtenus sur les jeux de données artificielles, nous avons dû changer notre méthode de travail.

Dans un premier temps, nous avons comparé les neuf distances disponibles dans le module DISS afin de choisir celles qui fournissent les résultats les plus convaincants pour les jeux de données artificielles.

Dans un second temps, nous avons cherché à savoir si les mauvais résultats obtenus par notre programme proviennent de la méthode employée, à savoir déterminer le nombre de classes à partir de la matrice de distance sans aucune modélisation, ou si ils proviennent des distances présentes dans le module DISS qui, pour rappel, ne sont pas les mêmes que celles utilisées par S. Delogne. Nous avons donc à cet effet, réalisé un nouveau programme et exploité les résultats ainsi obtenus.

Enfin, nous avons comparé les résultats des méthodes détermination du nombre de classes obtenus pour les trois programmes.

10.2 Analyse des distances présentes dans DISS

Comme nous l'avons signalé ci-dessus les résultats obtenus à l'aide du module DISS sont très médiocres. Nous allons donc dans cette section comparer les distances se trouvant dans le module DISS à savoir :

- U-1 : la mesure de dissimilarité de Gowda et Diday,
- U-2 : la première formulation de la mesure de dissimilarité d'Ichino et Yaguchi,
- U-3 : la mesure de dissimilarité d'Ichino et Yaguchi normalisée,
- U-4 : la mesure de dissimilarité d'Ichino et Yaguchi normalisée et pondérée,
- SO-1 : la mesure de dissimilarité de De Carvalho,
- SO-2 : la mesure de dissimilarité d'Ichino et Yaguchi étendue par De Carvalho,

– SO-3 : la première mesure de dissimilarité basée sur la description potentielle,
– SO-4 : la seconde mesure de dissimilarité basée sur la description potentielle et
– SO-5 : la mesure de dissimilarité basée sur la description potentielle normalisée.

Nous les comparerons à partir des résultats que nous avons obtenus sur deux des jeux artificiels de S. Delogne. Le premier jeu est celui contenant trois classes hyperspériques et le second est celui contenant deux classes allongées.

10.2.1 Données avec trois classes hypersphériques

Ce jeu de données est composé de trente objets symboliques décrits par deux variables de type intervalle. Ce jeu de données a été construit de sorte que l'on y retrouve trois classes naturelles hypersphériques bien séparées.

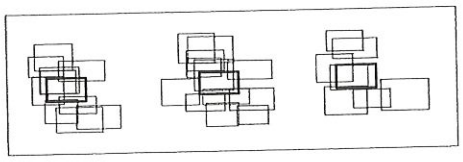


FIG. 10.1 – Données avec trois classes allongées

La distance U-1

Les méthodes de détermination du nombre de classes de Milligan et Cooper donnent les résultats suivants :

		M1	M2	M3	M4	M5
Saut minimum	-	2	×	×	×	×
Saut maximum	-	2	×	×	×	×
Centroïde	-	4	×	×	×	×
Ward	-	2	×	×	×	×

TAB. 10.1 – Résultats des méthodes de Milligan et Cooper pour les méthodes hiérarchiques : distance U-1

Nous remarquons que les méthodes de détermination du nombre de classes ne permettent pas de retrouver le bon nombre de classes quelle que soit la méthode de classification employée. De plus, si nous imposons trois classes, la partition retrouvée n'est pas la partition naturelle des données.

Les valeurs des indices pour différents nombres k de classes sont donnés pour la méthode du saut maximum dans le tableau 10.2.

Les indices associées aux méthodes de Duda et Hart (M2), du C-index (M3), de Gamma (M4) et de Beale (M5) ne sont pas significatifs et ne permettent pas de déterminer une structure. La méthode de Calinski et Harabaz ne permet pas de retrouver le bon nombre de classes.

Saut maximum	M1	M2	M3	M4	M5
k=8	3.13039	0.05868	0.08075	0.89987	0.33853
k=7	3.15187	0.21137	0.08874	0.88353	0.41709
k=6	3.15381	-0.11973	0.11716	0.82617	0.29993
k=5	3.31530	-0.07529	0.13372	0.78486	0.30085
k=4	3.34811	-0.45975	0.21625	0.60550	0.21934
k=3	3.39408	-0.84024	0.26772	0.44871	0.16212
k=2	**3.41027**	-0.40727	0.24832	0.44108	0.26238
k=1	-	-1.48965	-	-	0.11368

TAB. 10.2 – Valeurs des indices de Milligan et Cooper pour la méthode du saut maximum

Pour les partitions générées par SCLUST, comme on peut s'en rendre compte dans le tableau 10.4, les indices de (M3) et (M4) ne sont pas significatifs. La méthode de Calinski et Harabaz (M1) préconise elle un grand nombre de classes. Remarquons néanmoins que SCLUST retrouve la bonne classification en trois classes si on fixe le nombre de classes à trois.

		M1	M3	M4
SCLUST	+	5	×	×

TAB. 10.3 – Résultats des méthodes de Milligan et Cooper pour SCLUST : distance U-1

SCLUST	M1	M3	M4
k=8	1.66628	0.39571	0.49730
k=7	1.81415	0.33565	0.50693
k=6	1.73162	0.41667	0.41647
k=5	**1.91913**	0.38566	0.33880
k=4	1.73600	0.43216	0.21886
k=3	1.44850	0.49220	0.09897
k=2	1.59206	0.43051	0.12396
k=1	-	-	-

TAB. 10.4 – Valeurs des indices de Milligan et Cooper pour SCLUST

La distance U-2

Les méthodes de détermination du nombre de classes de Milligan et Cooper donnent les résultats suivants :

		M1	M2	M3	M4	M5
Saut minimum	+	3	3	3	3	3,×
Saut maximum	+	3	3	3	3	3,×
Centroïde	+	3	3	3	3	3,×
Ward	+	3	3	3	3	3,×

TAB. 10.5 – Résultats des méthodes de Milligan et Cooper pour les méthodes hiérarchiques : distance U-2

Les valeurs des indices de détermination du nombre de classes pour la méthode du centroïde se trouvent dans le tableau 10.6.

On remarque que les méthodes de détermination du nombre de classes déterminent dans tous les cas trois classes. Les indices de Duda et Hart (M2) et de Beale (M5) sont cependant peu significatifs dans le sens où les valeurs de rejet sugérées par Milligan et Cooper, eet Gordon ne sont pas atteintes, à savoir pour M2, 3.2 et $F_{p,(n-2)p}$ pour M5 qui dans ce cas vaut 3.15.

Lorsqu'une classification en trois classes est demandée pour les méthodes hiérarchiques, la bonne partition est retrouvée.

Centroïde	M1	M2	M3	M4	M5
k=8	19.14082	0.25668	0.01648	0.93694	0.44282
k=7	21.53843	0.04092	0.01586	0.94109	0.29877
k=6	23.59679	0.18455	0.01709	0.93979	0.42330
k=5	25.15470	0.26907	0.01339	0.95741	0.49855
k=4	27.25985	0.22051	0.00397	0.99240	0.48398
k=3	**38.10712**	**-0.07881**	**0.00002**	**0.99990**	**0.31658**
k=2	26.90079	1.48945	0.05422	0.82009	1.21840
k=1	-	1.21934	-	-	0.89669

TAB. 10.6 – Valeurs des indices de Milligan et Cooper pour la méthode du centroïde

Comme on le remarque dans les tableaux ci-dessous, la méthode SCLUST retrouve également les trois classes naturelles et la détermination du nombre de classes pour les trois méthodes de Milligan et Cooper est bien celle qui est attendue.

		M1	M3	M4
SCLUST	+	3	3	3

TAB. 10.7 – Résultats des méthodes de Milligan et Cooper pour SCLUST : distance U-2

SCLUST	M1	M3	M4
k=8	19.60876	0.02136	0.97511
k=7	22.44849	0.01249	0.97163
k=6	24.66426	0.01791	0.96800
k=5	27.21409	0.01955	0.95501
k=4	30.56215	0.01531	0.97478
k=3	**38.10712**	**0.00002**	**0.99990**
k=2	26.90079	0.05422	0.82009
k=1	-	-	-

TAB. 10.8 – Valeurs des indices de Milligan et Cooper pour SCLUST

La distance U-3

Les méthodes de détermination du nombre de classes de Milligan et Cooper donnent les résultats suivants :

		M1	M2	M3	M4	M5
Saut minimum	+	7,2	5	7	7	5
Saut maximum	-	7,2	7	7	7	7
Centroïde	+	7,2	7	7	7	7
Ward	-	7, 2	7	7	7	7

TAB. 10.9 – Résultats des méthodes de Milligan et Cooper pour les méthodes hiérarchiques : distance U-3

On remarque que la détermination de trois classes n'est jamais préconisée. Lorsque nous fixons le nombre de classes à trois, la partition naturelle en trois classes est retrouvée par deux des méthodes hiérarchiques, à savoir la méthode du saut minimum et celle du centroïde. La méthode SCLUST, comme le montre le tableau 10.10, retrouve la bonne partition en trois classes mais ne préconise pas la présence de trois classes dans les données.

		M1	M3	M4
SCLUST	+	7,2	6	6

TAB. 10.10 – Résultats des méthodes de Milligan et Cooper pour SCLUST : distance U-3

La distance U-4

Cette distance fournit identiquement les mêmes résultats que ceux obtenus en utilisant la distance précédente tant au niveau des méthodes hiérarchiques qu'au niveau de la méthode SCLUST.

La distance SO-1

Les méthodes de détermination du nombre de classes de Milligan et Cooper donnent les résultats suivants :

		M1	M2	M3	M4	M5
Saut minimum	-	6	×	×	×	×
Saut maximum	-	2	×	×	×	×
Centroïde	-	2	×	×	×	×
Ward	-	2	×	×	×	×

TAB. 10.11 – Résultats des méthodes de Milligan et Cooper pour les méthodes hiérarchiques : distance SO-1

On constate que les différentes méthodes hiérarchiques ne reconstituent pas la partition naturelle lorsque trois classes sont demandées et les méthodes de Milligan et Cooper ne préconisent dans aucun cas la présence de trois classes dans les données.

Les indices des différentes méthodes de détermination du nombre de classes pour la méthode hiérarchique de Ward sont donnés dans la table 10.12.

Une nouvelle fois, les différents méthodes de Milligan et Cooper ne retrouvent pas le bon nombre de classes pour les méthodes de classification hiérarchiques. Les indices de Duda et Hart (M2) et ceux de Beale (M5) ne permettent toujours pas de déterminer un nombre de classe avec conviction tout comme le C-index et l'indice Gamma.

La méthode SCLUST ne préconise pas le bon nombre de classes pour les méthodes du C-index et de Gamma. Seule la méthode de Calinski et Harabaz retrouve la bonne structure. L'utilisation de SCLUST permet de retrouver la bonne classification en trois classes comme nous pouvons le voir dans la table 10.13. De plus, les indices obtenus ne sont pas très "convaincants". En effet, dans le tableau 10.14, on remarque que la valeur maximale obtenue pour l'indice de Calinski-Harabaz (M1) n'est pas très marquée ; et que les valeurs obtenues pour le C-index (M3) et pour l'indice Gamma (M4) ne sont pas très proches de leur valeur idéale.

Centroïde	M1	M2	M3	M4	M5
k=10	2.30769	0.13509	0.06004	0.87575	0.37637
k=9	2.42357	0.50646	0.05828	0.87906	0.39735
k=8	2.52094	-0.08704	0.07605	0.84993	0.24304
k=7	2.67209	0.02591	0.08113	0.84320	0.29185
k=6	2.77833	-0.42973	0.12942	0.76449	0.18884
k=5	3.10830	-0.71008	0.14896	0.73498	0.12220
k=4	3.25862	-0.86040	0.21900	0.64561	0.14067
k=3	3.15751	-0.14609	0.22998	0.61542	0.32507
k=2	**3.47783**	-0.78172	0.21413	0.60510	0.17618
k=1	-	-1.47606	-	-	0.11593

TAB. 10.12 – Valeurs des indices de Milligan et Cooper pour la méthode du centroïde

		M1	M3	M4
SCLUST	+	3	×	×

TAB. 10.13 – Résultats des méthodes de Milligan et Cooper pour SCLUST : distance SO-1

SCLUST	M1	M3	M4
k=10	2.35817	0.15520	0.80416
k=9	2.40986	0.12943	0.81238
k=8	2.31242	0.18223	0.83100
k=7	2.60072	0.09148	0.83537
k=6	2.68178	0.14900	0.79240
k=5	2.88823	0.15089	0.73474
k=4	2.81256	0.20432	0.66888
k=3	**2.92127**	0.25452	0.58532
k=2	2.87506	0.29922	0.36637
k=1	-	-	-

TAB. 10.14 – Valeurs des indices de Milligan et Cooper pour la méthode SCLUST

La distance SO-2

Les méthodes de détermination du nombre de classes de Milligan et Cooper donnent les résultats suivants :

		M1	M2	M3	M4	M5
Saut minimum	+	3	×	9, 3	9, 3	×
Saut maximum	+	3	×	8, 3	8, 3	×
Centroïde	+	3	×	9, 3	9, 3	×
Ward	+	3	×	8, 3	8, 3	×

TAB. 10.15 – Résultats des méthodes de Milligan et Cooper pour les méthodes hiérarchiques : distance SO-2

Nous constatons que les quatre méthodes de classification hiérarchiques rerouvent bien les trois classes naturelles. La méthode de Calinski et Harabaz (M1) détecte dans tous les cas la présence de trois classes dans les données. Les méthodes basées sur les tests d'hypothèses (M2) et (M5) ne retrouvent pas de classes dans les données. Les indices du C-index et de Gamma suscitent une discussion quant à leur interprétation. Selon la méthode hiérarchique, on peut choisir de prendre soit neuf ou huit classes soit trois classes.

Les indices des différentes méthodes de détermination du nombre de classes pour la méthode hiérarchique du saut minimum sont donnés dans la table 10.15.

Pour la méthode du saut minimum, les indices trouvés pour les cinq méthodes de détermination du nombre de classes ne permettent pas de déterminer nettement un nombre de classes fixe excepté pour la méthode de Calinski et Harabaz (M1) qui possède un seul maximum. Les indices de détermination du nombre de classes de Duda et Hart (M2) et ceux de Beale (M5) ne sont pas significatifs et ne déterminent pas un nombre de classes k. Les autres indices suggèrent diverses possibilités pour déterminer le nombre de classe k. On pourrait, en effet, penser que la méthode du c-index (M3) préconise trois classes si nous considérons uniquement les résultats pour un petit nombre de classes.

La méthode SCLUST retrouve la bonne classification en trois classes et deux des méthodes de détermination du nombre de classes (M1) et (M3) préconisent la présence de trois classes dans les données ce qui n'est pas le cas de la méthode Gamma qui privilégie une structure des données en six classes.

Saut minimum	M1	M2	M3	M4	M5
k=10	3.87099	0.37740	0.03651	0.96716	0.48788
k=9	4.18766	-0.22713	**0.02908**	**0.97319**	0.18945
k=8	4.13516	0.05283	0.04923	0.94742	0.35775
k=7	4.40239	-0.45470	0.06276	0.92223	0.18091
k=6	4.85959	0.04352	0.05695	0.92798	0.33135
k=5	5.58712	-0.16686	0.05520	0.92661	0.24132
k=4	5.76693	0.07201	0.05219	0.91940	0.40459
k=3	**6.39148**	-0.21788	**0.03150**	**0.92290**	0.30791
k=2	6.17055	-0.49403	0.12085	0.65787	0.27359
k=1	-	-0.97821	-	-	0.20568

TAB. 10.16 – Valeurs des indices de Milligan et Cooper pour la méthode du saut minimum

		M1	M3	M4
SCLUST	-	3	3	6

TAB. 10.17 – Résultats des méthodes de Milligan et Cooper pour SCLUST : distance SO-2

SCLUST	M1	M3	M4
7	4.49533	0.05717	0.96945
6	5.18858	0.04870	**0.97761**
5	5.26994	0.07229	0.94247
4	5.42574	0.06591	0.93973
3	**6.39148**	**0.03150**	0.92290
2	6.17055	0.12085	0.65787
1	-	-	-

TAB. 10.18 – Valeurs des indices de Milligan et Cooper pour la méthode SCLUST

La distance SO-3

Les méthodes de détermination du nombre de classes de Milligan et Cooper donnent les résultats suivants :

		M1	M2	M3	M4	M5
Saut minimum	+	3	3	?,3	?,3	×,3
Saut maximum	+	3	3	?,3	?,3	×,3
Centroïde	+	3	3	?, 3	?, 3	×,3
Ward	+	3	3	?, 3	?, 3	×,3

TAB. 10.19 – Résultats des méthodes de Milligan et Cooper pour les méthodes hiérarchiques : distance SO-3

Dans le tableau ci-dessus, nous constatons que toutes les méthodes de classification hiérarchique partitionnent correctement le jeu de données en trois classes. Les méthodes de Calinski et Harabaz (M1), celle de Duda et Hart (M2) et celle de Beale (M5) recommandent la présence de trois classes dans les données. A nouveau les méthodes du c-index et de Gamma engendrent des discussions.

Les indices des différentes méthodes de détermination du nombre de classes pour la méthode hiérarchique de Ward sont donnés dans la table 10.20.

Ward	M1	M2	M3	M4	M5
k=10	22.65501	0.26902	0.00732	0.96607	0.45004
k=9	24.09994	0.80695	0.00816	0.96361	0.70057
k=8	24.93254	0.86814	0.00919	0.96145	0.96951
k=7	26.42907	0.55021	0.01259	0.95081	0.56488
k=6	29.42267	-0.20939	0.01584	0.94794	0.24743
k=5	28.61084	0.75730	0.02916	0.91204	0.82824
k=4	29.37152	1.04222	0.03103	0.90820	1.07123
k=3	**34.33463**	**0.55556**	**0.02164**	**0.93715**	**0.66007**
k=2	24.76706	1.38540	0.14264	0.70749	1.13256
k=1	-	1.07286	-	-	0.82557

TAB. 10.20 – Valeurs des indices de Milligan et Cooper pour la méthode de Ward

Le tableau 10.20 nous montre que la méthode de Calinski et Harabaz (M1) possède, comme pour la distance précédente, un seul maximum et détermine pour toutes les méthodes hiérarchiques trois classes. Les indices de détermination du nombre de classes de Duda et Hart (M2) et ceux de Beale (M5) sont peu significatifs mais permettent néan-

moins de déterminer un nombre de trois classes. Le C-index (M3) et l'indice Gamma (M4) engendrent une nouvelle fois une discussion quant à le détermination du nombre de classes.

		M1	M3	M4
SCLUST	+	3	?,7,3	?,6

TAB. 10.21 – Résultats des méthodes de Milligan et Cooper pour SCLUST : distance SO-3

En ce qui concerne SCLUST, la bonne classification est une nouvelle fois retrouvée mais les indices du C-index et de Gamma ont cependant un problème pour retrouver la bonne structure.

La distance SO-4

L'utilisation de cette distance donne exactement les mêmes résultats que lorsque nous utilisons la distance SO-3 dans le module de départ.

La distance SO-5

Cette distance fournit pratiquement les mêmes résultats que ceux obtenus lors de l'utilisation de SO-2. La seule différence se situe au niveau du C-index (M3) qui dans cas n'éveille pas diverses interprétations. Pour les quatres méthodes hiérarchiques, le C-index (M3) préconise uniquement la présence de trois classes dans les données.

10.2.2 Données avec deux classes allongées

Ce jeu de données est composé de trente objets symboliques décrits par deux variables de type intervalle. Les données ont été réparties en deux classes allongées.

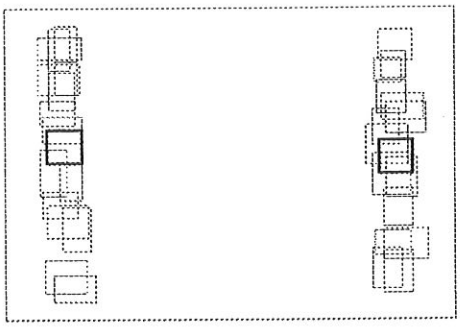


FIG. 10.2 – Données avec deux classes allongées

La distance U-1

Les méthodes de détermination du nombre de classes de Milligan et Cooper donnent les résultats suivants :

		M1	M2	M3	M4	M5
Saut minimum	+	2	×	×	×	×
Saut maximum	+	2	×	×	×	×
Centroïde	+	2	×	×	×	×
Ward	+	2	×	×	×	×

TAB. 10.22 – Résultats des méthodes de Milligan et Cooper pour les méthodes hiérarchiques : distance U-1

Premièrement, nous constatons que la partition en deux classes donnée par les méthodes hiérarchiques est la partition naturelle des données. La méthode de détermination du nombre de classes de Calinski et Harabaz (M1), du C-index et de Gamma préconisent deux classes pour les quatre méthodes de classification hiérarchiques. La méthode de Duda et Hart (M2) et celle de Beale (M5) ne permettent en aucun cas de déterminer un nombre de classe étant donné les valeurs des indices qui ne sont pas significatives du tout.

Les valeurs des indices pour différents nombres k de classes sont donnés pour la méthode de Ward dans le tableau 10.23.

Ward	M1	M2	M3	M4	M5
k=8	2.71863	-0.12742	0.12346	0.82363	0.22686
k=7	2.68969	-0.16186	0.15900	0.76322	0.28331
k=6	2.85010	-0.22514	0.14598	0.78203	0.24152
k=5	3.03957	-0.68939	0.15733	0.74782	0.15355
k=4	3.23577	-0.07563	0.18138	0.69877	0.30071
k=3	3.48383	-0.49611	0.17831	0.70187	0.21868
k=2	**4.16976**	-1.06754	0.17506	0.58242	0.13304
k=1	-	-1.34017	-	-	0.13899

TAB. 10.23 – Valeurs des indices de Milligan et Cooper pour la méthode de Ward

nt

La méthode SCLUST classifie correctement les données et les méthodes de détermination du nombre de classes préconise deux classes pour cette même méthode comme le montre la table 10.24.

		M1	M3	M4
SCLUST	+	2	2	2

TAB. 10.24 – Résultats des méthodes de Milligan et Cooper pour SCLUST : distance U-1

La distance U-2

Les méthodes de détermination du nombre de classes de Milligan et Cooper pour les quatre méthodes hiérarchiques donnent les résultats présentés dans le tableau 10.25.

		M1	M2	M3	M4	M5
Saut minimum	+	2	2	2	2	×,2
Saut maximum	+	2	2	2	2	×,2
Centroïde	+	2	2	2	2	×,2
Ward	+	2	2	2	2	×,2

TAB. 10.25 – Résultats des méthodes de Milligan et Cooper pour les méthodes hiérarchiques : distance U-2

Les deux classes naturelles sont retrouvées par les quatre méthodes hiérarchiques. Toutes les méthodes de détermination retrouvent le bon nombre de classes excepté la méthode de Beale qui posséde des indices peu significatifs.

Les valeurs des indices pour différents nombres k de classes sont données dans la table 10.26 pour la méthode du centroïde.

Centroïde	M1	M2	M3	M4	M5
k=8	33.94333	0.73771	0.00674	0.98224	0.80427
k=7	33.02702	0.51328	0.01352	0.96774	0.63442
k=6	32.61910	0.77555	0.01389	0.96672	0.84495
k=5	33.70710	1.00656	0.01473	0.96561	1.12523
k=4	38.32487	0.80790	0.01640	0.96403	0.88652
k=3	33.55588	1.04980	0.04081	0.92947	0.96694
k=2	**40.86748**	**1.23296**	**0.00000**	**1.00000**	**1.12025**
k=1	-	1.95396	-	-	1.36225

TAB. 10.26 – Valeurs des indices de Milligan et Cooper pour la méthode du centroïde

Touts les indices sont également très significatifs et aucun doute n'est permis quant à la présence de deux classes dans les données.

Les méthodes de détermination du nombre de classes pour la méthode SCLUST fournissent les résultats suivants :

		M1	M3	M4
SCLUST	+	2	2	2

TAB. 10.27 – Résultats des méthodes de Milligan et Cooper pour SCLUST : distance U-2

Ce tableau montre que SCLUST retrouve la bonne partition des données en deux classes et que les trois méthodes de Milligan et Cooper fonctionnent correctement.

Les indices sont à nouveau très significatifs et sont donnés dans le tableau 10.28.

SCLUST	M1	M3	M4
k=5	33.94636	0.02223	0.96739
k=4	37.14461	0.01896	0.97649
k=3	33.55588	0.04081	0.92947
k=2	**40.86748**	**0.00000**	**1.00000**
k=1	-	-	-

TAB. 10.28 – Valeurs des indices de Milligan et Cooper pour SCLUST

La distance U-3

Les méthodes de détermination du nombre de classes de Milligan et Cooper pour les quatre méthodes hiérarchiques donnent les résultats suivants :

		M1	M2	M3	M4	M5
Saut minimum	+	2	4, 2	2	2	7,4
Saut maximum	+	4	3	2	2	6,3
Centroïde	+	4	3	2	2	6,3
Ward	+	4	3	2	2	6,3

TAB. 10.29 – Résultats des méthodes de Milligan et Cooper pour les méthodes hiérarchiques : distance U-3

Comme pour la distance U-3, les méthodes hiérarchiques retrouvent les classes naturelles. Les méthodes du C-index (M3) et de Gamma (M4) détectent deux classes dans les données pour les quatre méthodes hiérarchiques. Les méthodes de Calinski-Harabaz (M1), de Duda et Hart (M2) et de Beale (M5) présentent quant à elles des problèmes.

Les valeurs des indices pour différents nombres k de classes sont données dans la table 10.30 pour la méthode du saut maximum.

Saut maximum	M1	M2	M3	M4	M5
k=8	28.14938	0.84136	0.01124	0.97619	0.93249
k=7	26.33915	0.68655	0.01781	0.96721	0.76668
k=6	25.51433	0.84669	0.01870	0.96635	**0.91357**
k=5	25.86189	1.06854	0.02003	0.96522	1.21499
k=4	**28.91452**	0.85627	0.02248	0.96354	0.93874
k=3	23.82123	**1.08111**	0.05922	0.92929	**0.99145**
k=2	26.40968	1.25697	**0.00210**	**0.98561**	1.14227
k=1	-	1.18664	-	-	0.88032

TAB. 10.30 – Valeurs des indices de Milligan et Cooper pour la méthode du saut maximum

Pour la méthode SCLUST, les méthodes de détermination du nombre de classes donnent les résultats suivants :

		M1	M3	M4
SCLUST	+	4	2	2

TAB. 10.31 – Résultats des méthodes de Milligan et Cooper pour SCLUST : distance U-3

Ce tableau montre que SCLUST retrouve la bonne partition des données en deux classes. Les méthodes du C-index (M3) et de Gamma (M4) retrouvent le bon nombre de classes. Néanmoins, la méthode de Calinski et Harabaz (M1) pose problème quant à la détermination du nombre de classes.

Les valeurs des indices pour les partitions produites par SCLUST sont présentées dans le tableau ci-après.

SCLUST	M1	M3	M4
k=5	26.04128	0.03014	0.96726
k=4	**27.93288**	0.02611	0.97649
k=3	23.82123	0.05922	0.92929
k=2	26.40968	**0.00210**	**0.98561**
k=1	-	-	-

TAB. 10.32 – Valeurs des indices de Milligan et Cooper pour SCLUST

La distance U-4

Les résultats de détermination du nombre de classes en utilisant la distance U-4 sont les mêmes que ceux obtenus lorsqu'on utilise la distance U-3 tant au niveau des méthodes hiérarchiques qu'au niveau de la méthode non-hiérarchique SCLUST.

La distance SO-1

Les méthodes de détermination du nombre de classes de Milligan et Cooper donnent les résultats suivants :

		M1	M2	M3	M4	M5
Saut minimum	+	2	×	2	2	×
Saut maximum	+	2	×	2	2	×
Centroïde	+	2	×	2	2	×
Ward	+	2	×	2	2	×

TAB. 10.33 – Résultats des méthodes de Milligan et Cooper pour les méthodes hiérarchiques : distance SO-1

Une nouvelle fois, les deux classes sont retrouvées par les quatre méthodes de classification hiérarchiques. Seules les méthodes basées sur les tests d'hypothèse ne conseillent pas la présence de deux classes dans les données.

Nous présentons les indices de Milligan et Cooper pour la méthode du saut minimum dans le tableau suivant :

Ward	M1	M2	M3	M4	M5
k=8	2.75647	0.64395	0.05901	0.78410	0.65369
k=7	2.82355	0.20681	0.06728	0.76417	0.38302
k=6	2.81617	-0.10182	0.06625	0.76593	0.30715
k=5	2.90742	-0.69414	0.07671	0.75166	0.16321
k=4	3.11698	-0.78216	0.07970	0.76184	0.15051
k=3	3.80583	-1.05825	0.06907	0.79698	0.10449
k=2	**5.86499**	-1.02938	**0.03704**	**0.90421**	0.11079
k=1	-	-1.03072	-	-	0.19550

TAB. 10.34 – Valeurs des indices de Milligan et Cooper pour la méthode de Ward

Comme nous pouvons le voir les méthodes de Calinski et Harabaz (M1), du C-index (M3) et de Gamma (M4) retrouvent le bon nombre de classes. Les indices relatifs aux méthodes de Duda et Hart (M2) et de Beale (M5) ne possède aucun indice significatif.

Si nous analysons les résultats donnés par les méthodes de Milligan et Cooper appliquées aux partitions générées par SCLUST, nous obtenons

		M1	M3	M4
SCLUST	+	2	2	2

TAB. 10.35 – Résultats des méthodes de Milligan et Cooper pour SCLUST : distance SO-1

SCLUST	M1	M3	M4
k=5	2.94677	0.08420	0.78003
k=4	3.26513	0.08612	0.76366
k=3	3.97470	0.06553	0.80177
k=2	**5.86499**	**0.03704**	**0.90421**
k=1	-	-	-

TAB. 10.36 – Valeurs des indices de Milligan et Cooper pour la méthode SCLUST

On remarque que les différents indices préconisent la présence de deux classes dans les données comme le prouve les indices présentés dans la table 10.36 et en plus SCLUST retrouve la bonne classification.

La distance SO-2

Quand on utilise cette distance dans le module DISS, les différentes méthodes de Milligan et Cooper donnent les mêmes résultats que lorsqu'on utilise la distance SO-1. Les indices de détermination du nombre de classes sont cependant plus significatifs excepté pour les méthodes basées sur les tests d'hypothèse.

La distance SO-3

Les méthodes de détermination du nombre de classes de Milligan et Cooper donnent les résultats présentés dans la table suivante :

		M1	M2	M3	M4	M5
Saut minimum	+	2	2	2	2	2
Saut maximum	+	2	2	2	2	2
Centroïde	+	2	2	2	2	2
Ward	+	2	2	2	2	2

TAB. 10.37 – Résultats des méthodes de Milligan et Cooper pour les méthodes hiérarchiques : distance SO-3

Nous remarquons que les cinq méthodes de détermination du nombre de classes retrouvent le bon nombre de classes dans les données pour les quatre méthodes de classification hiérarchiques. Il en est de même pour la méthode non-hiérarchique SCLUST. Les partitions engendrées par les différentes méthodes de classification sont également exactes.

Nous présentons les indices de Milligan et Cooper pour la méthode du centroïde dans le tableau suivant :

Centroïde	M1	M2	M3	M4	M5
k=8	79.06336	0.70529	0.00238	0.96997	0.76846
k=7	79.69286	0.51454	0.00330	0.96020	0.63530
k=6	82.09311	0.58542	0.00316	0.96284	0.68658
k=5	88.09284	0.90667	0.00310	0.96404	0.99693
k=4	101.11857	0.73912	0.00416	0.95567	0.81769
k=3	100.86979	0.93338	0.01023	0.92735	0.88122
k=2	**139.65704**	**1.15045**	**0.00058**	**0.99915**	**1.04813**
k=1	-	3.65549	-	-	4.65523

TAB. 10.38 – Valeurs des indices de Milligan et Cooper pour la méthode du centroïde

Une nouvelle fois, ces indices sont très significatifs et ne pose aucun problème quant à leur interprétation.

		M1	M3	M4
SCLUST	+	2	2	2

TAB. 10.39 – Résultats des méthodes de Milligan et Cooper pour SCLUST : distance SO-3

La distance SO-4

Cette distance fournit les mêmes résultats que lors de l'utilisation de la distance SO-3 pour la détermination du nombre de classes et de classification.

La distance SO-5

L'utilisation de cette distance fournit des résultats similaires à ceux obtenus lors de l'utilisation de la distance SO-1.

10.2.3 Conclusion

Pour les quatre méthodes hiérarchiques, il est très clair que les distances U-2, SO-3 et SO-4 fournissent de meilleurs résultats que les six autres distances pour les deux jeux de

données. Les indices sont en effet très significatifs lors de l'utilisation de U-2. L'utilisation de SO-3 ou SO-4 dans le module DISS engendre néanmoins plusieurs interprétations. L'utilisation de certaines distances telles que U-1,U-3, U-4 ne permettent pas de détecter la structure naturelle des deux jeux de données. Il en est de même pour les distances S0-1, SO-2 et SO-5 pour les données de formes hypersphériques. Pour ces mêmes distances, lors de la détermination du nombre de classes pour le jeu de données avec deux classes allongées, les indices de Duda et Hart et de Beale ne sont jamais significatifs.

La méthode SCLUST retrouve la bonne classification des jeux de données.
A nouveau, seule la distance U-2 préconise le bon nombre de classes pour les données réparties en trois classes hypersphériques pour toutes les méthodes de détermination du nombre de classes.
Pour les données réparties en deux classes allongées, SCLUST recommande quasiment dans tous les cas une structure des données en deux classes sauf lorsque les distances U-3 et U-4 sont associées à l'indice de Calinski et Harabaz.

Nous comparerons donc uniquement les résultats obtenus à partir de la distance U-2 avec ceux obtenus par S. Delogne. En effet, si nous n'obtenons pas de bons résultats sur des données artificielles avec certaines distances, il nous semble dès lors difficile de les utiliser sur des ensembles de données réelles.

10.3 Nouveau programme

Comme, nous l'avons signalé à plusieurs reprises, le méthode que nous avons adaptée est assez médiocre et ne fournit pas toujours les résultats attendus. Mais d'où vient le problème ? Est-ce le fait de ne pas avoir employé une modélisation comme l'avait fait S. Delogne qui obtenait de très bons résultats ? Ou le problème vient-il des distances implémentées dans le module DISS ?

Nous savons que la distance de Hausdorff est celle qui donnait les meilleurs résultats dans l'étude réalisée par S. Delogne. Nous avons donc réalisé un programme qui calcule la matrice de distance avec cette distance de Hausdorff mais n'utilise aucune modélisation des données. Nous avons ensuite réalisé le même raisonnement que dans notre première méthode.

Ce programme se trouve sur cd à la fin de ce rapport et se nomme intervalle-Hausdorff.

10.3.1 Résultats obtenus pour les données avec trois classes hypersphériques

Les méthodes de détermination du nombre de classes de Milligan et Cooper donnent les résultats suivants :

		M1	M2	M3	M4	M5
Saut minimum	+	3	3	3	3	3,×
Saut maximum	+	3	3	3	3	3,×
Centroïde	+	3	3	3	3	3, ×
Ward	+	3	3	3	3	3,×

TAB. 10.40 – Résultats des méthodes de Milligan et Cooper pour les méthodes hiérarchiques : distance de Hausdorff

Les trois classes naturelles sont retrouvées pour chaque méthode de classification hiérarchique. Toutes les méthodes de détermination du nombre de classes ont retrouvé le bon nombre de classes.

Les valeurs des indices pour différents nombres k de classes sont donnés pour la méthode du saut minimum dans le tableau 10.41.

Saut maximum	M1	M2	M3	M4	M5
k=8	18.50450	0.76619	0.01071	0.95920	0.83731
k=7	20.37930	0.73519	0.01207	0.95445	0.60735
k=6	22.50420	0.22272	0.01389	0.94938	0.42342
k=5	24.69066	0.16215	0.01984	0.93714	0.42651
k=4	28.10970	0.33308	0.01576	0.95584	0.53231
k=3	**34.84315**	**0.16576**	**0.00010**	**0.99990**	**0.45908**
k=2	25.68926	1.36700	0.05394	0.82081	1.11810
k=1	-	1.13760	-	-	0.85631

TAB. 10.41 – Valeurs des indices de Milligan et Cooper pour la méthode du saut maximum

Tous les indices sont significatifs, sauf ceux des méthodes basées sur des tests d'hypothèse pour lesquels les valeurs de rejet classiques ne sont pas atteintes.

Les résultats des méthodes de détermination du nombre de classes de Calinski et Harabaz (M1), du C-index (M3) et Gamma (M4) appliquées aux partitions générées par SCLUST se trouvent dans le tableau 10.42. Similairement aux quatre méthodes de classification hiérarchiques, SCLUST retrouve la bonne partition des données en trois classes.

		M1	M3	M4
SCLUST	+	3	3	3

TAB. 10.42 – Résultats des méthodes de Milligan et Cooper pour SCLUST : distance de Hausdorff

Les valeurs des indices des méthodes de Milligan et Cooper sont reprises dans le tableau 10.43.

SCLUST	M1	M3	M4
k=5	24.43108	0.01982	0.95682
k=4	27.84492	0.01543	0.97614
k=3	**34.84315**	**0.00010**	**0.99990**
k=2	25.68926	0.05394	0.82081
k=1	-	-	-

TAB. 10.43 – Valeurs des indices de Milligan et Cooper pour SCLUST : distance de Hausdorff

10.3.2 Résultats obtenus pour les données avec deux classes allongées

Les méthodes de détermination du nombre de classes de Milligan et Cooper appliquées aux quatre hiérarchies de partitions produites par les quatre méthodes hiérarchiques donnent les résultats présentés dans le tableau 10.44.

		M1	M2	M3	M4	M5
Saut minimum	+	2	2	2	2	2,×
Saut maximum	+	2	2	2	2	2,×
Centroïde	+	2	2	2	2	2, ×
Ward	+	2	2	2	2	2,×

TAB. 10.44 – Résultats des méthodes de Milligan et Cooper pour les méthodes hiérarchiques : distance de Hausdorff

Nous constatons tout d'abord que les quatre méthodes de classification hiérarchiques retrouvent bien les deux classes naturelles. Les cinq méthodes de Milligan et Cooper détectent deux classes dans les données pour les méthodes hiérarchiques.

Les valeurs des indices des cinq méthodes de détermination du nombre de classes de Milligan et Cooper appliquées à la hiérarchie de partitions générée par la méthode de Ward

sont donnés dans le tableau 10.45.

Ward	M1	M2	M3	M4	M5
k=8	29.14058	0.80204	0.00948	0.97217	0.88118
k=7	28.60464	0.49817	0.01361	0.96590	0.62397
k=6	28.80778	0.65072	0.01403	0.96613	0.73730
k=5	30.32577	0.87251	0.01445	0.96522	0.95706
k=4	34.89261	0.67923	0.01586	0.96391	0.76240
k=3	31.51492	0.94042	0.04005	0.93017	0.88617
k=2	**39.11191**	**1.10946**	**0.00000**	**1.00000**	**1.01421**
k=1	-	1.87842	-	-	1.30373

TAB. 10.45 – Valeurs des indices de Milligan et Cooper pour la méthode de Ward

Les résultats fournis par la méthode SCLUST se trouvent dans le tableau 10.46.

		M1	M3	M4
SCLUST	+	2	2	2

TAB. 10.46 – Résultats des méthodes de Milligan et Cooper pour SCLUST : distance de Hausdorff

Tout comme les quatre méthodes de classification hiérarchiques, SCLUST retrouve la bonne partition des données en deux classes naturelles. Les trois méthodes de Milligan et Cooper vont dans le sens des résultats obtenus précédemment.

Les valeurs des indices pour les partitions produites par SCLUST sont présentés dans le tableau 10.47.

SCLUST	M1	M3	M4
k=5	30.45609	0.02059	0.98310
k=4	34.89261	0.01586	0.98486
k=3	31.51492	0.04005	0.93017
k=2	**39.11191**	**0.00000**	**1.00000**
k=1	-	-	-

TAB. 10.47 – Valeurs des indices de Milligan et Cooper pour la méthode SCLUST

10.4 Comparaison des résultats sur diverses jeux de données

Dans cette section, nous comparons la modélisation milieu-longueur de Stéphane Delogne, la méthode réalisée avec le module DISS et le programme créé avec la distance de Hausdorff. Nous rappelerons les résultats obtenus pour chaque méthode. Pour rappel, pour la méthode que nous avons réalisée en utilisant le module DISS, nous nous limitons à la distance fournissant les meilleurs résultats, à savoir U-2.

10.4.1 Données avec trois classes hypersphériques

Il s'agit du même jeu de données que celui étudié précédemment.

Résultats obtenus avec la modélisation milieu-longueur

Les méthodes de détermination du nombre de classes de Milligan et Cooper donnaient les résultats suivants :

		M1	M2	M3	M4	M5
Saut minimum	+	3	3	3	3	3
Saut maximum	+	3	3	3	3	3
Centroïde	+	3	3	3	3	3
Ward	+	3	3	3	3	3

TAB. 10.48 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux quatre méthodes hiérarchiques

S. Delogne retrouvait les trois classes naturelles pour chaque méthode de classification hiérarchique et toutes les méthodes de détermination du nombre de classes retrouvaient le bon nombre de classes.

Les valeurs des indices pour différents nombres k de classes sont données pour la méthode du centroïde dans le tableau 10.49. Tous les indices étaient significatifs.

Les trois méthodes de détermination du nombre de classes applicables aux partitions générées par SCLUST retrouvaient également le bon nombre de classes et SCLUST retrouvait la bonne classification en trois classes (tableau 10.50). Les indices étaient encore très significatifs.

Centroïde	M1	M2	M3	M4	M5
k=8	147.46524	0.93037	0.00214	0.96108	1.04810
k=7	147.36367	0.61831	0.00287	0.95356	0.70219
k=6	151.08969	0.75366	0.00355	0.95240	0.83174
k=5	142.02988	1.29472	0.00409	0.95895	1.38609
k=4	138.96187	0.99391	0.00168	0.99245	0.97659
k=3	**189.95941**	**0.41215**	**0.00002**	**0.99990**	**0.56737**
k=2	67.04706	2.93795	0.05495	0.81996	4.24084
k=1	-	2.74934	-	-	2.23490

TAB. 10.49 – Valeurs des indices de Milligan et Cooper pour la méthode du centroïde

		M1	M3	M4
SCLUST	+	3	3	3

TAB. 10.50 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à SCLUST

Les différents indices sont à nouveau très significatifs comme nous pouvons le constater :

SCLUST	M1	M3	M4
k=5	156.35205	0.00574	0.95018
k=4	165.37531	0.00527	0.96158
k=3	**189.95941**	**0.00002**	**0.99990**
k=2	67.04706	0.05495	0.81996
k=1	-	-	-

TAB. 10.51 – Valeurs des indices de Milligan et Cooper pour SCLUST

Résultats obtenus avec le module DISS

L'utilisation de la distance U-2 dans le module DISS permet sans problème de retrouver la bonne classification des données en trois classes et les cinq méthodes de Milligan et Cooper permettent de retrouver le bon nombre de classes tant au niveau des méthodes hiérarchiques qu"au niveau de la méthode non hiérachique.

Résultats obtenus en utilisant la distance de Hausdorff

Nous obtenons les mêmes résultats que ceux obtenus en utilisant le module DISS.

Conclusion

Aucune des méthodes ne semble meilleure qu'une autre si ce n'est la méthode utilisant la modélisation milieu-longueur. En effet, quand on consulte les indices de détermination du nombre de classes pour les méthodes hiérarchiques dans la tableau 10.49, on constate que tous les indices sont très significatifs même ceux des méthodes de Duda et Hart (M2) et de Beale (M1). Pour ces dernières, dans le cas de l'utilisation du module DISS et pour la distance de Hausdorff, les valeurs de rejet classiques ne sont pas atteintes.

10.4.2 Données avec deux classes allongées

Résultats obtenus pour la modélisation milieu-longueur

Les méthodes de détermination du nombre de classes de Milligan et Cooper appliquées aux quatre méthodes hiérarchiques donnaient les résultats suivants :

		M1	M2	M3	M4	M5
Saut minimum	+	2	2	2	2	2
Saut maximum	+	2	2	2	2	2
Centroïde	+	2	2	2	2	2
Ward	+	2	2	2	2	2

TAB. 10.52 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux quatre méthodes hiérarchiques

Toutes les méthodes de Milligan et Cooper donnaient le bon nombre de classes. De plus, toutes les méthodes hiérarchiques retrouvaient la bonne partition des données en deux classes.

Les valeurs des indices des méthodes de Milligan et Cooper sont données dans le tableau suivant pour la méthode de classification de Ward.

Ward	M1	M2	M3	M4	M5
k=8	298.60394	1.11538	0.00143	0.98023	1.42009
k=7	244.49843	1.17224	0.00365	0.96607	1.32139
k=6	214.10616	1.51514	0.00427	0.96599	2.04243
k=5	199.56491	1.70226	0.00494	0.96509	3.13591
k=4	208.93498	1.45955	0.00600	0.96360	2.07302
k=3	118.11523	2.24511	0.02666	0.93003	2.82678
k=2	**126.24044**	**2.38480**	**0.00000**	**1.00000**	**3.32345**
k=1	-	3.55231	-	-	4.20801

TAB. 10.53 – Valeurs des indices de Milligan et Cooper pour la méthode de Ward

Tous ces indices sont significatifs. Aucun doute n'est donc possible quant à la présence de deux classes dans les données.

Les méthodes de détermination du nombre de classes de Milligan et Cooper appliquées aux partitions générées par SCLUST fournissaient les résultats :

		M1	M3	M4
SCLUST	+	2	2	2

TAB. 10.54 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à SCLUST

Cela signifie que SCLUST retrouve la bonne partition des données en deux classes et que les trois méthodes de Milligan et Cooper ont l'air de bien fonctionner.

Les indices sont également très significatifs comme nous pouvons le constater dans le tableau 10.55.

SCLUST	M1	M3	M4
k=5	199.56491	0.00494	0.97754
k=4	208.93498	0.00600	0.97754
k=3	118.11523	0.02666	0.93003
k=2	**126.24044**	**0.00000**	**1.00000**
k=1	-	-	-

TAB. 10.55 – Valeurs des indices de Milligan et Cooper pour SCLUST

Résultats obtenus avec le module DISS

Tout comme pour les données réparties en trois classes hypersphériques, les différentes méthodes de classification retrouvent la bonne partition des données. Quant à la détermination du nombre de classes, les cinq méthodes sont unanimes ; elles préconisent toutes la présence de deux classes dans les données. Une nouvelle fois les indices des méthodes de Duda et Hart (M2) et de Beale (M5) ne sont pas significatifs.

Résultats obtenus avec la distance de Hausdorff

Les résultats obtenus sont en tout point similaires à ceux obtenus lors de l'utilisation du module DISS.

Conclusion

Toutes les méthodes fournissent une nouvelle fois des résultats semblables. A nouveau, la modélisation milieu-longueur permet d'obtenir des indices très significatifs pour les cinq méthodes de Milligan et Cooper.

10.4.3 Données sans structure

Ce jeu de données artificiel a été construit de sorte qu'on ne puisse trouver aucune structure dans les données. Il est composé de cinquante objets symboliques décrits par trois variables de type intervalle. Nous rappelerons les résultats que S. Delogne avait obtenus. Nous étudierons ensuite les résultats que nous obtenons en utilisant le module DISS et le distance U-2 et ceux obtenus lorsque nous calculons la matrice de distance grâce à la distance de Hausdorff. Nous comparerons enfin l'entièreté des résultats.

Nous tenons à faire remarquer tout comme S. Delogne l'avait fait que les indices des méthodes de détermination du nombre de classes de Milligan et Cooper sont relativement difficiles à interpréter, notamment en raison de leur faible signification. Delogne avait néanmoins essayé d'attribuer un nombre de classes à chaque méthode de Milligan et Cooper appliquée aux méthodes hiérarchiques dans un premier temps et à SCLUST ensuite. Nous ferons de même.

Résultats obtenus pour la modélisation milieu-longueur

Les résultats que Delogne avait fournis pour les méthodes hiérarchiques sont les suivants :

	M1	M2	M3	M4	M5
Saut minimum	2	×	5	×	×
Saut maximum	3,7	×	×	×	8
Centroïde	2	×	×	×	×
Ward	3	×	×	×	×

TAB. 10.56 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux quatre méthodes hiérarchiques

Les différentes méthodes de Milligan et Cooper retrouvent rarement des classes dans les données.

Seule la méthode de Calinski et Harabasz (M1) a tendance à retrouver deux voire trois classes. Les indices relatifs aux méthodes de Duda et Hart (M2) et de Beale (M5) ne permettent pas de déterminer le nombre de classes dans la mesure où les valeurs de rejet

classiques ne sont jamais atteintes. Les méthodes du C-index (M3) et Gamma (M4) présentent quant à elles, respectivement, des indices croissants et décroissants en fonction de l'agrégation des objets. Il est donc très difficile de tirer une conclusion autre que l'absence de structure.

Nous présentons les indices des méthodes de Milligan et Cooper pour la méthode du saut maximum dans le tableau 10.57. Un simple examen de ce tableau nous permet de constater que les valeurs des différents indices ne nous suggèrent pas un nombre de classes bien précis.

Saut maximum	M1	M2	M3	M4	M5
k=8	6.17785	1.14438	0.11837	0.65605	1.16836
k=7	7.00370	-0.42807	0.13554	0.61246	0.14011
k=6	6.66799	0.64498	0.17179	0.53170	0.69683
k=5	6.96280	0.30942	0.18863	0.48872	0.50232
k=4	6.98372	0.43910	0.21642	0.44823	0.57761
k=3	**8.23114**	-0.27264	0.24893	0.40151	0.29783
k=2	8.09276	-0.18087	0.39524	0.20413	0.36223
k=1	-	-0.68865	-	-	0.27173

TAB. 10.57 – Valeurs des indices de Milligan et Cooper pour la méthode du saut maximum

Les résultats donnés par ces méthodes appliquées aux partitions générées par SCLUST sont donnés dans le tableau 10.58.

	M1	M3	M4
SCLUST	2	1	6

TAB. 10.58 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à SCLUST

Les indices sont tout aussi variés que précédemment. La méthode de Calinski et Harabasz (M1) propose une fois de plus deux classes, les méthode du C-index (M3) et Gamma (M4) suggèrent, respectivement, l'absence de structure et la présence de six classes.

Les valeurs des différents indices des méthodes de Milligan et Cooper appliquées aux partitions générées par SCLUST sont présentées dans le tableau 10.59.

SCLUST	M1	M3	M4
k=10	5.90538	0.13212	0.60988
k=9	6.34816	0.12698	0.58209
k=8	6.82539	0.13414	0.58506
k=7	6.94743	0.14621	0.56865
k=6	7.37663	0.15042	**0.58984**
k=5	7.66971	0.18110	0.55118
k=4	7.95596	0.21082	0.50564
k=3	9.22466	0.23397	0.48812
k=2	**10.69524**	0.29502	0.44111
k=1	-	-	-

TAB. 10.59 – Valeurs des indices de Milligan et Cooper pour SCLUST

Résultats obtenus avec le module DISS

Les méthodes de détermination du nombre de classes de Milligan et Cooper pour les quatre méthodes hiérarchiques donnent les résultats suivants :

	M1	M2	M3	M4	M5
Saut minimum	2	×	×	×	×
Saut maximum	2	×	×	×	×
Centroïde	10	×	×	×	×
Ward	2	×	×	×	×

TAB. 10.60 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux quatre méthodes hiérarchiques

Les différentes méthodes de Milligan et Cooper retrouvent rarement des classes dans les données.

Comme lors de l'étude de Stéphane Delogne, seule la méthode de Calinski et Harabasz (M1) a tendance à retrouver deux classes excepté pour la méthode du centroïde pour laquelle M1 préconise un nombre élevé de classes dans les données. Les indices relatifs aux méthodes de Duda et Hart (M2), du C-index (M3),de Gamma (M4) et de Beale (M5) ne permettent pas de déterminer le nombre de classes.

Nous présentons les indices des méthodes de Milligan et Cooper pour la méthode du centroïde dans le tableau 10.61. Un simple examen de ce tableau nous permet de constater que les valeurs des différents indices ne nous suggèrent pas un nombre de classes bien précis.

Centroïde	M1	M2	M3	M4	M5
k=10	4.64040	0.36260	0.07965	0.78695	0.49142
k=9	4.94156	0.12712	0.08234	0.77883	0.36001
k=8	4.65836	-0.22847	0.11516	0.69513	0.32528
k=7	4.64216	0.22188	0.12494	0.67028	0.46330
k=6	4.47169	0.43168	0.13907	0.63506	0.57384
k=5	3.82607	-0.52434	0.18406	0.56982	0.27591
k=4	2.19422	-0.64089	0.17437	0.75704	0.27676
k=3	2.25556	-1.69832	0.18650	0.77080	0.06913
k=2	**2.40990**	-1.71045	0.19404	0.78464	0.07017
k=1	-	-1.66659	-	-	0.08092

TAB. 10.61 – Valeurs des indices de Milligan et Cooper pour la méthode du centroïde

Les résultats donnés par ces méthodes appliquées aux partitions générées par SCLUST fournissent :

	M1	M3	M4
SCLUST	2	×	×

TAB. 10.62 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à SCLUST

Les indices sont tout aussi semblables que précédemment. La méthode de Calinski et Harabasz (M1) propose une fois de plus deux classes, les méthode du C-index (M3) et Gamma (M4) suggèrent, toutes les deux, l'absence de structure.

Les valeurs des différents indices des méthodes de Milligan et Cooper appliquées aux partitions générées par SCLUST sont présentées dans le tableau 10.63.

Résultats obtenus avec la distance de Hausdorff

Les indices sont une nouvelle semblable à ceux obtenus pour la méthode DISS. Pour cette raison, nous ne fournissons que les résultats obtenus pour la détermination du nombre de classes dans un premier temps pour les quatre méthodes de classification hiérarchique et dans un second temps pour la méthode non hiérarchique SCLUST.

Les méthodes de détermination du nombre de classes de Milligan et Cooper pour les quatre méthodes hiérarchiques donnent les résultats présentés dans le tableau 10.64.

SCLUST	M1	M3	M4
k=10	4.31798	0.12481	0.68867
k=9	4.84934	0.12227	0.67970
k=8	3.97759	0.19611	0.57062
k=7	5.04947	0.14163	0.57940
k=6	6.11962	0.11980	0.64027
k=5	4.38768	0.24250	0.56725
k=4	5.43396	0.21449	0.55257
k=3	6.18136	0.25087	0.52372
k=2	**8.05821**	0.25948	0.50547
k=1	-	-	-

TAB. 10.63 – Valeurs des indices de Milligan et Cooper pour SCLUST

	M1	M2	M3	M4	M5
Saut minimum	2	×	×	3	×
Saut maximum	2	×	×	×	×
Centroïde	7,4	×	×	×	×
Ward	2	×	×	×	×

TAB. 10.64 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux quatre méthodes hiérarchiques

Les méthodes de détermination du nombre de classes pour la méthode SCLUST donnent les résultats suivants :

	M1	M3	M4
SCLUST	2	×	×

TAB. 10.65 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à SCLUST

Conclusion

Il est très difficile de tirer des conclusions pour ce jeu de données. En effet, étant donné qu'il est sans structure, on ne peut pas vérifier l'efficacité des différentes méthodes de détermination du nombre de classes. Néanmoins, toutes les méthodes se valent dans le sens où elles fournissent toutes des résultats similaires.

10.4.4 Données avec deux classes emboîtées

Ce jeu de données est composé de vingt objets symboliques décrits par une seule variable de type intervalle. Ces données ont été simulées à l'aide du logiciel statistique S-PLUS de manière à mettre en évidence deux classes d'objets. Le centre des intervalles est compris entre 0 et 10 tandis que la longueur est comprise entre 1 et 2 pour l'une des deux classes, entre 11 et 12 pour l'autre.

Résultats obtenus pour la modélisation milieu-longueur

Les méthodes de détermination du nombre de classes de Milligan et Cooper appliquées aux hiérarchies de partitions produites par les quatre méthodes hiérarchiques donnaient les résultats suivants :

		M1	M2	M3	M4	M5
Saut minimum	+	5,2	2	2	2	6
Saut maximum	+	4	2	2	2	8
Centroïde	+	4	2	2	2	7
Ward	+	4	2	2	2	8

TAB. 10.66 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux quatre méthodes hiérarchiques

Nous constatons tout d'abord que les quatre méthodes de classification hiérarchiques retrouvent bien les deux classes naturelles. Les méthodes de Duda et Hart (M2), du C-index (M3) et Gamma (M4) détectent deux classes dans les données pour les quatre méthodes hiérarchiques. Les méthodes de Calinski et Harabasz (M1) et de Beale (M5) présentent quant à elles quelques problèmes, surtout cette dernière qui a tendance à retrouver un grand nombre de classes.

Les valeurs des indices des cinq méthodes de détermination du nombre de classes de Milligan et Cooper appliquées à la hiérarchie de partitions générée par la méthode du saut minimum sont données dans le tableau 10.67.

Les partitions générées par SCLUST donnaient les résultats présentés dans le tableau 10.68.

Tout comme les quatre méthodes de classification hiérarchiques, SCLUST retrouve la bonne partition des données en deux classes. Pour ce qui est du nombre de classes, les méthodes du C-index (M3) et Gamma (M4) vont dans le sens des résultats obtenus précédemment, à savoir la présence de deux classes naturelles. La méthode de Calinski et Harabasz (M1) est encore discordante avec les autres méthodes dans la mesure où elle

Saut minimum	M1	M2	M3	M4	M5
k=8	272.56390	0.73008	0.00117	0.98235	1.23168
k=7	178.10013	0.31709	0.00170	0.98339	0.64550
k=6	136.46791	0.80954	0.00205	0.98234	**1.41124**
k=5	**153.52362**	0.89848	0.00157	0.98733	3.09742
k=4	56.96081	0.98872	0.03134	0.90846	1.46431
k=3	56.84995	0.77254	0.01221	0.97078	1.05560
k=2	**88.00246**	**-1.46910**	**0.00000**	**1.00000**	0.15430
k=1	-	1.40643	-	-	1.51728

TAB. 10.67 – Valeurs des indices de Milligan et Cooper pour la méthode du saut minimum

		M1	M3	M4
SCLUST	+	4	2	2

TAB. 10.68 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à SCLUST

propose trois classes.

Les valeurs des indices pour les partitions produites par SCLUST sont présentées dans le tableau 10.69.

SCLUST	M1	M3	M4
k=5	153.52362	0.00157	0.98733
k=4	**147.36065**	0.00426	0.97540
k=3	98.57734	0.01591	0.95397
k=2	88.00246	**0.00000**	**1.00000**
k=1	-	-	-

TAB. 10.69 – Valeurs des indices de Milligan et Cooper pour SCLUST

Résultats obtenus avec le module DISS

Les méthodes de détermination du nombre de classes de Milligan et Cooper appliquées aux hiérarchies de partitions produites par les quatre méthodes hiérarchiques fournissent les résultats suivants :

		M1	M2	M3	M4	M5
Saut minimum	+	4	1	4	4	1,6
Saut maximum	+	4	1	?	6	1,8
Centroïde	+	4	1	5	5	1,7
Ward	+	4	1	?	6	1,8

TAB. 10.70 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux quatre méthodes hiérarchiques : DISS

Nous constatons qu'aucune des cinq méthodes de détermination du nombre de classses ne retrouvent le bon nombre de classes dans les données.

Les valeurs des indices des cinq méthodes de détermination du nombre de classes de Milligan et Cooper appliquées à la hiérarchie de partitions générée par la méthode du centroïde sont données dans le tableau 10.71.

Centroïde	M1	M2	M3	M4	M5
k=10	17.41706	0.23220	0.00908	0.98295	0.48512
k=9	15.24016	0.12282	0.01449	0.97415	0.41429
k=8	16.07327	0.64734	0.01090	0.98235	0.92857
k=7	12.82512	-0.20619	0.01156	0.98418	**0.37659**
k=6	13.20857	0.67212	0.01148	0.98492	1.00433
k=5	12.91775	0.12274	**0.00873**	**0.98766**	0.46939
k=4	**13.83340**	-0.08975	0.02165	0.97571	0.36386
k=3	11.53036	-0.41236	0.06026	0.88429	0.35830
k=2	9.81898	-0.29482	0.07349	0.71546	0.39332
k=1	-	-2.06098	-	-	0.16929

TAB. 10.71 – Valeurs des indices de Milligan et Cooper pour la méthode du centroïde

Les partitions générées par SCLUST fournissent quant à elles les résultats présentés dans le tableau 10.72.

SCLUST retrouve la bonne partition des données en deux classes. Pour ce qui est de la détermination du nombre de classes, les résultats sont tout aussi mauvais que ceux obtenus pour les méthodes hiérarchiques.

		M1	M3	M4
SCLUST	+	4	?	4

TAB. 10.72 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à SCLUST

Les valeurs des indices pour les partitions produites par SCLUST sont présentées dans le tableau 10.73.

SCLUST	M1	M3	M4
k=7	14.92623	0.02091	0.97077
k=6	12.75272	0.06197	0.92208
k=5	12.68434	0.02856	0.96398
k=4	**13.83340**	0.02165	**0.97571**
k=3	11.53036	0.06026	0.88429
k=2	9.81898	0.07349	0.71546
k=1	-	-	-

TAB. 10.73 – Valeurs des indices de Milligan et Cooper pour SCLUST

Résultats obtenus avec la distance de Hausdorff

Les méthodes de détermination du nombre de classes de Milligan et Cooper appliquées aux quatre méthodes hiérarchiques donnent les résultats suivants :

		M1	M2	M3	M4	M5
Saut minimum	+	2	×	5	5	×
Saut maximum	+	4, 2	×	6	6	×
Centroïde	+	4, 2	×	5	5	×
Ward	+	4, 2	×	6	6	×

TAB. 10.74 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux quatre méthodes hiérarchiques : Hausdorff

Toutes les méthodes de classification hiérarchiques retrouvent la bonne classification lorsque deux classes sont demandées. Néanmoins, les méthodes de Milligan et Cooper ne sont pas très convaincantes quant à la présence de deux classes dans les données excepté pour la méthode de Calinski et Harabaz (M1).

Les valeurs des indices des méthodes de Milligan et Cooper sont données dans le tableau suivant pour la méthode de classification du saut minimum.

Saut minimum	M1	M2	M3	M4	M5
k=8	14.06045	-0.41494	0.01254	0.96875	0.30903
k=7	15.98160	0.60045	0.00623	0.98299	0.80952
k=6	14.86529	0.06717	0.00805	0.97957	0.43834
k=5	16.97696	0.54479	**0.00609**	**0.98467**	0.69858
k=4	11.58373	-0.14412	0.06404	0.88359	0.43256
k=3	11.64582	-0.53934	0.02924	0.93793	0.32432
k=2	**18.84556**	-2.08218	0.02831	0.92422	0.08599
k=1	-	-0.90926	-	-	0.32492

TAB. 10.75 – Valeurs des indices de Milligan et Cooper pour la méthode du saut minimum

Les méthodes basées sur les tests d'hypothèse possédent encore des indices peu significatifs, elles ne retrouvent aucune structure dans les données. Les méthodes du C-index et de Gamma ont des indices fort proches l'un de l'autre. Il est donc difficile d'affirmer la présence de tel ou tel autre nombre de classes. Nous avons néanmoins fait un choix. Seule la méthode de Calinski et Harabaz fournit une valeur significative et ne pose aucun problème pour l'interprétation.

Les méthodes de détermination du nombre de classes de Milligan et Cooper appliquées aux partitions générées par SCLUST fournissent les résultats ci-après.

		M1	M3	M4
SCLUST	+	4, 2	?	7, ?

TAB. 10.76 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à SCLUST

Cela signifie que SCLUST retrouve la bonne partition des données en deux classes et que les trois méthodes de Milligan et Cooper ont l'air de bien fonctionner.

Les indices sont peu significatifs comme nous pouvons le constater dans le tableau 10.77.

Conclusion

Au niveau de la classification, les trois méthodes étudiées donnent de bons résultas dans le sens où les partitions fournies par toutes les méthodes de classification hiérarchiques ou non lorsque deux classes sont demandées sont équivalentes à la partition naturelle des données.

SCLUST	M1	M3	M4
k=8	16.62845	0.01448	0.97998
k=7	18.34241	0.01333	0.99427
k=6	16.53748	0.03435	0.96044
k=5	16.97281	0.01559	0.96364
k=4	18.90072	0.01221	0.97354
k=3	17.51486	0.03269	0.94068
k=2	18.84556	0.02831	0.92422
k=1	-	-	-

TAB. 10.77 – Valeurs des indices de Milligan et Cooper pour SCLUST

Sans nul doute, la méthode basée sur la modélisation milieu-longueur fournit de bien meilleurs résultats en ce qui concerne les méthodes de détermination du nombre de classes. En effet, trois des cinq méthodes retrouvent le bon nombre de classes.

10.4.5 Données basées sur les formes de Breiman [1]

Cet ensemble de données provient d'un problème de reconnaissance de formes d'ondes issu de l'étude de L. Breiman [13].

Nous disposons de 30 objets symboliques décrits par 21 variables de type intervalle répartis en trois classes.

Les données sont initialement composées de 3000 points décrits par 21 variables quantitatives continues et répartis en trois classes de 1000 points chacune.

Le modèle est basé sur les formes d'ondes h_1, h_2 et h_3 tel que :

- $h_1(i) = \max\{6 - |i-7|, 0\}$;
- $h_2(i) = h_1(i+4)$ et
- $h_3(i) = h_1(i+8)$.

L'ensemble d'apprentissage est composé de trois formes d'ondes décrites par 21 variables $(X_j;\ j = 1, \ldots, 21)$ données par la formule :

- $X_j = U\, h_1(j) + (1-U)\, h_2(j) + \varepsilon_j,\ j = 1, \ldots, 21$ pour les ondes du premier groupe ;
- $X_j = U\, h_1(j) + (1-U)\, h_3(j) + \varepsilon_j,\ j = 1, \ldots, 21$ pour les ondes du second groupe ;
- $X_j = U\, h_2(j) + (1-U)\, h_3(j) + \varepsilon_j,\ j = 1, \ldots, 21$ pour les ondes du troisième groupe ;

où

- U est la variable aléatoire uniforme sur l'intervalle $[0, 1]$;

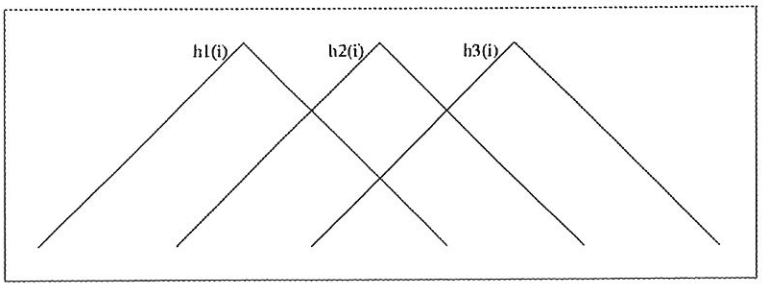


FIG. 10.3 – Les trois formes d'ondes du modèle de Breiman

- $\{\varepsilon_j \ ; \ j = 1, \ldots, 21\}$ sont des variables aléatoires gaussiennes indépendantes, de moyenne 0 et de variance 1 ;
- h_1, h_2 et h_3 sont les fonctions des formes d'ondes.

La matrice de données classiques (x_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, 21$) est donc générée par :

- $x_{ij} = u_i\, h_1(j) + (1 - u_i)\, h_2(j) + \varepsilon_{ij}$, $j = 1, \ldots, 21$ pour le point i du groupe des ondes du premier groupe ;
- $x_{ij} = u_i\, h_1(j) + (1 - u_i)\, h_3(j) + \varepsilon_{ij}$, $j = 1, \ldots, 21$ pour le point i du groupe des ondes du second groupe ;
- $x_{ij} = u_i\, h_2(j) + (1 - u_i)\, h_3(j) + \varepsilon_{ij}$, $j = 1, \ldots, 21$ pour le point i du groupe des ondes du troisième groupe.

La projection des données dans le plan principal est donnée par la figure 10.4.

Trois classes semblent se dégager de cette projection des données dans le plan principal, avec peut-être des ponts entre les classes.

A l'aide du module DB2SO du logiciel SODAS, qui permet de générer un ensemble de données symboliques à partir de données classiques, 30 objets symboliques décrits par 21 variables de type intervalle ont été construits. On peut y retrouver trois classes comptant chacune 10 objets. Chaque classe d'objets symboliques correspond à une classe d'objets classiques.

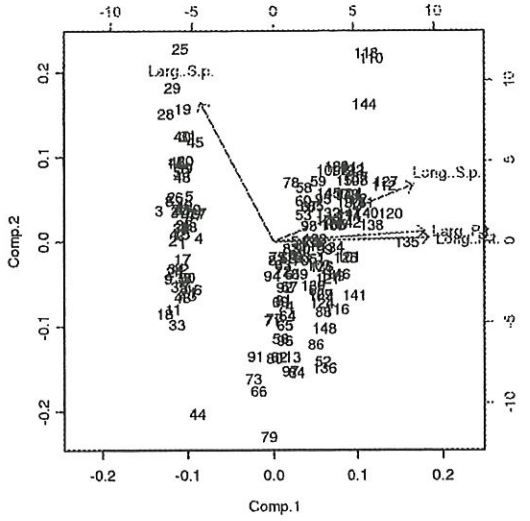


FIG. 10.4 – Projection des données dans le plan principal

Résultats obtenus pour la modélisation milieu-longueur

Les résultats obtenus pour la modélisation milieu-longueur pour les méthodes de détermination du nombre de classes de Milligan et Cooper aux hiérarchies de partitions produites par les quatre méthodes hiérarchiques sont repris dans le tableau suivant :

		M1	M2	M3	M4	M5
Saut minimum	-	4	4	?	?	4
Saut maximum	-	3	4	?	?	4
Centroïde	-	3	4	?	?	4
Ward	-	3	4	?	?	4

TAB. 10.78 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux quatre méthodes hiérarchiques

Aucune méthode de classification ne retrouve la bonne partition des données en trois classes. En effet, chacune des méthodes a tendance à regrouper des objets symboliques de deux classes différentes dans la même classe. Les méthodes de détermination du nombre de classes n'ont donc pas été appliquées à une bonne classification.

En conséquence, les nombres de classes proposés ne sont pas unanimes en faveur d'une structure des données en trois classes. La méthode de Calinski et Harabasz (M1) retrouve trois classes pour la plupart des méthodes de classification. Les méthodes de Duda et Hart

(M2) et de Beale (M5) ont quant à elles tendance à suggérer l'existence de quatre classes pour toutes les méthodes hiérarchiques. Les méthodes qui rencontrent le plus de problèmes sont celles du C-index (M3) et Gamma (M4). En effet, l'interprétation des valeurs des indices relatifs à ces méthodes nous incite à conclure, soit à l'existence de dix classes, soit à l'absence de structure. Remarquons qu'un nombre de dix classes est fort élevé pour une base de données ne comportant que trente objets.

Les valeurs des différents indices des méthodes de Milligan et Cooper sont présentées pour la méthode de Ward :

Ward	M1	M2	M3	M4	M5
k=10	11.26786	2.27816	**0.00752**	**0.96938**	1.20652
k=9	11.86979	3.20881	0.01062	0.95642	1.76844
k=8	12.62642	2.52520	0.01136	0.95118	1.66267
k=7	13.38024	3.51203	0.01473	0.93688	2.40985
k=6	14.73487	2.95095	0.01790	0.92432	1.94599
k=5	15.82149	2.75337	0.03077	0.89011	2.06168
k=4	17.70035	**3.52864**	0.03258	0.89293	**2.88538**
k=3	**20.16070**	4.73021	0.05220	0.84805	4.62904
k=2	19.49813	5.58004	0.14328	0.68787	5.39876
k=1	-	6.88179	-	-	6.54517

TAB. 10.79 – Valeurs des indices de Milligan et Cooper pour la méthode de Ward

En ce qui concerne les résultats donnés par les méthodes de Milligan et Cooper sur les partitions générées par SCLUST, les nombres de classes proposés sont encore discordants (tableau 10.80).

		M1	M3	M4
SCLUST	-	2	6	5

TAB. 10.80 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à SCLUST

Ici aussi, la partition des trente objets symboliques en trois classes n'est toujours pas celle attendue.

Les indices des méthodes de Milligan et Cooper prennent les valeurs suivantes :

SCLUST	M1	M3	M4
k=8	11.70199	0.01829	0.95391
k=7	12.72180	0.02200	0.95239
k=6	14.65995	**0.01723**	0.95710
k=5	16.53562	0.01977	**0.96801**
k=4	18.52572	0.02540	0.94705
k=3	22.44191	0.03483	0.95258
k=2	**22.73835**	0.09586	0.89281
k=1	-	-	-

TAB. 10.81 – Valeurs des indices de Milligan et Cooper pour SCLUST

Résultats obtenus avec le module DISS

Les résultats que nous obtenons pour les méthodes de détermination du nombre de classes de Milligan et Cooper aux hiérarchies de partitions produites par les quatre méthodes hiérarchiques sont les suivants :

		M1	M2	M3	M4	M5
Saut minimum	-	2	2	×	×	2
Saut maximum	-	3	3	3	3	3
Centroïde	-	3	3	?	?	3
Ward	-	3	3	?	?	3

TAB. 10.82 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux quatre méthodes hiérarchiques

Les différentes méthodes de classification ne retrouvent pas la bonne partition des données en trois classes comme dans le cas de la modélisation milieu-longueur. Les méthodes de détermination du nombre de classes n'ont donc toujours pas été appliquées à une bonne classification.

Comme ci-dessus, les nombres de classes proposés ne sont pas unanimes en faveur d'une structure des données en trois classes. La méthode de Calinski et Harabasz (M1) retrouve trois classes pour la plupart des méthodes de classification. Les méthodes de Duda et Hart (M2) et de Beale (M5) proposent également la plupart du temps une structure en trois classes. Les méthodes qui rencontrent le plus de problèmes sont une nouvelle fois celles du C-index (M3) et Gamma (M4). En effet, nous pouvons interpréter les valeurs obtenues de

diverses façons. Pour la méthode du saut minimum, les valeurs sont très peu significatives dans le sens où, pour le C-index, les plus petites valeurs sont proches de 0.1 et pour Gamma, les valeurs les plus élevées sont proches 0.9. Pour les autres méthodes de classification, les indices sont fort semblables.

Les valeurs des différents indices des méthodes de Milligan et Cooper sont présentées pour la méthode du centroïde.

Centroïde	M1	M2	M3	M4	M5
k=10	6.29882	2.66105	0.01531	0.96387	1.26738
k=9	6.43001	2.50187	0.02661	0.93902	1.64233
k=8	7.06604	3.07700	0.02686	0.93825	1.15632
k=7	7.38847	2.83773	0.03231	0.92513	1.95133
k=6	7.63971	3.05358	0.04355	0.90234	2.25431
k=5	8.15674	3.31560	0.05797	0.86946	2.34235
k=4	9.75128	2.25592	0.05748	0.87128	1.57325
k=3	**9.81153**	**3.55494**	0.13376	0.71287	**2.96233**
k=2	8.82992	3.83559	0.22604	0.52963	3.14190
k=1	-	3.79095	-	-	2.96405

TAB. 10.83 – Valeurs des indices de Milligan et Cooper pour la méthode de Ward

Pour les partitions générées par SCLUST, les méthodes de détermination du nombre de classes donnent les résultats suivants :

		M1	M3	M4
SCLUST	-	3	?	?

TAB. 10.84 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à SCLUST

Une nouvelle fois, la partition des objets en trois classes n'est toujours pas la partition naturelle. Le nombre de classes préconisé pour les différentes méthodes de Milligan et Cooper n'est pas très clair. Comme précédemment, pour la méthode du C-index et de Gamma, les valeurs sont fort proches l'une de l'autre et ne permettent pas une interprétation aisée.

Les indices des méthodes de Milligan et Cooper prennent les valeurs reprises dans le tableau 10.85.

SCLUST	M1	M3	M4
k=10	6.05237	0.02463	0.94304
k=9	6.43186	0.02474	0.95993
k=8	6.64178	0.04383	0.95170
k=7	7.45032	0.03210	0.96373
k=6	8.43295	0.02930	0.97228
k=5	8.94370	0.04509	0.95809
k=4	10.58966	0.03557	0.96138
k=3	**12.71150**	0.04378	0.96668
k=2	12.71117	0.12006	0.83688
k=1	-	-	-

TAB. 10.85 – Valeurs des indices de Milligan et Cooper pour SCLUST

Résultats obtenus avec la distance de Hausdorff

Lorsque les indices de Milligan et Cooper sont calculés avec la matrice de distance utilisant la distance de Hausdorff, nous obtenons les résultats présentés dans le tableau 10.86 pour les quatre méthodes de classification hiérarchiques.

		M1	M2	M3	M4	M5
Saut minimum	-	3	3	?	?	3
Saut maximum	-	2	2	?	?	2
Centroïde	-	2	2	?	?	2
Ward	-	2	2	?	?	2

TAB. 10.86 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux quatre méthodes hiérarchiques

Une fois de plus, la partition naturelle n'est pas retrouvée quand nous sollicitons trois classes dans les données. Et donc, les indices de Milligan et Cooper sont "faussés".

Trois des méthodes de détermination du nombre de classes revendiquent une structure en deux classes dans les données excepté pour la méthode du saut minimum pour laquelle trois classes sont recommandées. L'interprétation des indices C-index et Gamma est encore malaisée.

Les valeurs des différents indices des méthodes de Milligan et Cooper sont présentées pour la méthode du saut minimum dans le tableau 10.87.

saut minimum	M1	M2	M3	M4	M5
k=8	4.35576	2.36998	0.04960	0.89285	1.66932
k=7	4.73649	1.23848	0.05251	0.88675	0.86654
k=6	5.16761	1.34672	0.06384	0.86286	0.96768
k=5	5.70831	1.43002	0.08483	0.81564	1.04962
k=4	6.92240	1.75061	0.07083	0.84772	1.18117
k=3	9.51401	1.58758	0.06682	0.85673	1.06520
k=2	9.21389	4.34175	0.15973	0.65449	3.81534
k=1	-	3.93294	-	-	3.09294

TAB. 10.87 – Valeurs des indices de Milligan et Cooper pour la méthode de Ward

Les méthodes de détermination du nombre de classes associées à la méthode non hiérarchique SCLUST produisent les résultats ci-après :

		M1	M3	M4
SCLUST	-	2	?	?

TAB. 10.88 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à SCLUST

La bonne structure des données n'est pas retrouvée et les valeurs proposées pour les méthodes de Milligan et Cooper ne sont toujours pas très précises.

SCLUST	M1	M3	M4
k=8	5.28046	0.04453	0.95609
k=7	5.85709	0.03605	0.96504
k=6	6.54896	0.03296	0.97591
k=5	7.01871	0.04766	0.95997
k=4	8.21986	0.03896	0.96065
k=3	9.97211	0.04882	0.96290
k=2	**10.61411**	0.11239	0.85226
k=1	-	-	-

TAB. 10.89 – Valeurs des indices de Milligan et Cooper pour SCLUST

Conclusion

Il est très difficile de tirer des conclusions sur les résultats fournis par les différents programmes. En effet, aucun programme ne permet de retrouver ne fusse que pour une méthode de classification la bonne partition des données en trois classes.

Les méthodes de détermination du nombre de classes n'ayant pas été appliquées à des classifications correctes, les nombres de classes proposés par celles-ci sont donc érronés.

10.4.6 Températures de villes chinoises

Nous allons ici étudier un jeu de données réel. La base de données répertorie les températures enregistrées dans soixante stations chinoises durant l'année 1988.
Une variable de type intervalle a été combinée à chaque mois de cette année. Les soixante stations sont par conséquent décrites par douze variables. Pour une station et un mois fixe, les bornes inférieure et supérieure de chaque intervalle correspondent, respectivement, aux températures minimale et maximale mesurée pendant ce mois et dans cette station.

Nous appliquons uniquement les méthodes de détermination du nombre de classes à ce jeu de données. En effet, les données sont réelles et nous ne connaissons pas le nombre de classes composé par celles-ci. Diverses interprétations de ces données ont déjà été fournies pour deux et cinq classes.

Résultats obtenus pour la modélisation milieu-longueur

Les méthodes de détermination du nombre de classes de Milligan et Cooper appliquées aux méthodes hiérarchiques donnaient les résultats suivants pour la modélisation milieu-longueur :

	M1	M2	M3	M4	M5
Saut minimum	4	4	7	4	4
Saut maximum	2	7	?	?	7
Centroïde	2	8	?	?	6
Ward	2	7	?	?	6

TAB. 10.90 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux quatre méthodes hiérarchiques

On constate que la méthode de Calinski et Harabasz (M1) a tendance à retrouver deux classes dans les données. Les autres méthodes sont complètement discordantes. En effet, les

méthodes de Duda et Hart (M2) et de Beale (M5) retrouvent des nombres de classes différents pour pratiquement toutes les méthodes de classification. Les méthodes du C-index (M3) et Gamma (M4) suggèrent quant à elles l'absence de structure dans les données dans le sens où toutes les valeurs des indices sont proches les unes des autres. De ces résultats, il devient très laborieux de tirer des conclusions.

Les valeurs des indices sur les hiérarchies de partitions produites par la méthode du saut maximum sont reprises dans le tableau 10.91.

Saut maximum	M1	M2	M3	M4	M5
k=10	66.97910	3.35843	0.00710	0.94317	3.28357
k=9	68.76579	3.49386	0.00813	0.93705	3.63604
k=8	71.52922	2.33948	0.01046	0.92269	2.00180
k=7	74.55677	**3.15177**	0.01203	0.91400	**2.85955**
k=6	71.40533	4.39313	0.01766	0.88751	6.02929
k=5	70.03495	6.32987	0.02792	0.83522	10.33830
k=4	69.23276	4.56283	0.03766	0.83376	3.74062
k=3	73.50859	4.93518	0.05416	0.80128	5.29241
k=2	**86.52993**	6.55223	0.07487	0.80923	5.55238
k=1	-	10.72136	-	-	10.52002

TAB. 10.91 – Valeurs des indices de Milligan et Cooper pour la méthode du saut maximum

En ce qui concerne les partitions produites par SCLUST, nous obtenons :

	M1	M3	M4
SCLUST	2	?	?

TAB. 10.92 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à SCLUST

Comme précédemment, la méthode de Calinski et Harabasz (M1) suggère l'existence de deux classes dans les données. Les méthodes du C-index (M3) et Gamma (M4) proposent une nouvelle fois des valeurs peu significatives pour déterminer le nombre de classes. Si on veut absolument citer un nombre de classes, il faut en choisir six.

Les indices des méthodes de Milligan et Cooper associés aux partitions générées par SCLUST sont fournis par le tableau 10.93.

SCLUST	M1	M3	M4
k=7	71.48622	0.01471	0.90847
k=6	79.37365	**0.01432**	**0.91144**
k=5	75.03564	0.02479	0.86042
k=4	75.91534	0.02971	0.89517
k=3	72.15914	0.06218	0.89911
k=2	**89.13224**	0.07557	0.85281
k=1	-	-	-

TAB. 10.93 – Valeurs des indices de Milligan et Cooper pour SCLUST

Résultats obtenus avec le module DISS

Les méthodes de détermination du nombre de classes de Milligan et Cooper appliquées aux méthodes hiérarchiques fournissent, lorsque nous utilisons dans un premier temps le module DISS, les résultats suivants :

	M1	M2	M3	M4	M5
Saut minimum	2	8	×	×	8
Saut maximum	2	3	?	?	2
Centroïde	2	8, 4	?	?	2
Ward	2	3	?	?	2

TAB. 10.94 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux quatre méthodes hiérarchiques

On constate que la méthode de Calinski et Harabasz (M1) et celle de Beale (M5) ont tendance à retrouver deux classes dans les données. La méthode de Duda et Hart (M2) propose des résultats assez mitigés. Elle préconise parfois huit parfois trois classes dans les données. Les méthodes du C-index (M3) et Gamma (M4) suggèrent diverses interprétations. Pour la méthode du saut minimum, les indices ne sont pas assez significatifs et feraient penser à un jeu de données sans structure. Les autres méthodes de classification possèdent des valeurs assez semblables pour les différents nombres de classes et il nous est impossible de décider.

Les valeurs des indices sur les hiérarchies de partitions produites par la méthode du saut minimum sont reprises dans le tableau 10.95.

Saut minimum	M1	M2	M3	M4	M5
k=10	6.43025	0.31233	0.11126	0.71242	0.43538
k=9	7.11313	0.23006	0.10739	0.72404	0.39654
k=8	8.09170	**2.34767**	0.10638	0.72705	**1.14351**
k=7	2.91194	5.74622	0.34075	0.40049	4.04445
k=6	3.43557	2.68126	0.33887	0.40572	1.58252
k=5	3.78721	-0.36258	0.34865	0.40318	0.23719
k=4	4.66518	-0.60076	0.33444	0.44460	0.14501
k=3	5.32641	-0.03809	0.25235	0.63420	0.37551
k=2	**9.28603**	-0.57859	0.18484	0.76188	0.16360
k=1	-	1.66930	-	-	1.12897

TAB. 10.95 – Valeurs des indices de Milligan et Cooper pour la méthode du saut minimum

Pour les partitions produites par SCLUST, nous obtenons les résultats suivant :

	M1	M3	M4
SCLUST	2	?	?

TAB. 10.96 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à SCLUST

Comme précédemment, la méthode de Calinski et Harabasz (M1) suggère l'existence de deux classes dans les données. Les méthodes du C-index (M3) et Gamma (M4) proposent une nouvelle fois des valeurs peu significatives pour déterminer le nombre de classes. Si nous voulons absolument choisir un nombre de classes, il faut citer le nombre six.

Les indices des méthodes de Milligan et Cooper associés aux partitions générées par SCLUST sont fournis par le tableau 10.97.

Résultats obtenus avec la distance de Hausdorff

Le programme utilisant la distance de Hausdorff fournit les résultats repris dans le tableau 10.98 pour les méthodes de détermination du nombre de classes de Milligan et Cooper appliquées aux méthodes hiérarchiques.

Trois des méthodes sont assez unanimes quant à la présence de deux classes dans les données. Comme pour le programme utilisant le module DISS, les méthodes du C-index et Gamma ne possèdent pas des indices très significatfifs et ne permettent pas une interprétation aisée.

SCLUST	M1	M3	M4
k=7	23.46960	0.01972	0.93695
k=6	24.84682	0.02727	0.91496
k=5	26.91028	0.03913	0.88073
k=4	27.44589	0.04318	0.88575
k=3	29.20836	0.07940	0.80437
k=2	35.58353	0.10220	0.80312
k=1	-	-	-

TAB. 10.97 – Valeurs des indices de Milligan et Cooper pour SCLUST

	M1	M2	M3	M4	M5
Saut minimum	4	8	×	×	8
Saut maximum	2	2	?	?	2
Centroïde	2	2	?	?	2
Ward	2	2	?	?	2

TAB. 10.98 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux quatre méthodes hiérarchiques

Les valeurs des indices sur les hiérarchies de partitions produites par la méthode de Ward sont reprises dans le tableau 10.99.

En ce qui concerne les partitions produites par SCLUST, nous obtenons les résultats présentés dans le tableau 10.100.

Comme précédemment, la méthode de Calinski et Harabasz (M1) suggère l'existence de deux classes dans les données et les deux autres méthodes posent le même problème que précédemment.

Les indices des méthodes de Milligan et Cooper associés aux partitions générées par SCLUST sont fournis par le tableau 10.101.

Conclusion

La grande variété des nombres de classes proposés par les différentes méthodes pour ce jeu de données réel vient du fait que les hiérarchies de partitions et partitions produites par les méthodes de classification sont bien souvent différentes. Il est alors très délicat de conclure quant à l'existence d'un certain nombre de classes plutôt qu'à un autre.

Ward	M1	M2	M3	M4	M5
k=10	16.88738	2.30264	0.01429	0.95948	1.50372
k=9	17.83038	2.06979	0.01711	0.95043	1.71840
k=8	18.77038	2.31786	0.01899	0.94469	1.96842
k=7	20.29463	1.44475	0.02697	0.92088	1.19131
k=6	21.74336	3.01426	0.03309	0.90278	2.90158
k=5	23.11424	3.53686	0.04889	0.85634	3.31229
k=4	24.13693	3.00204	0.05069	0.85676	2.28711
k=3	26.63665	3.38544	0.08011	0.79151	2.89225
k=2	**33.67928**	**3.76691**	0.08555	0.80250	**2.68404**
k=1	-	6.17580	-	-	4.09462

TAB. 10.99 – Valeurs des indices de Milligan et Cooper pour la méthode de Ward

	M1	M3	M4
SCLUST	2	?	?

TAB. 10.100 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à SCLUST

Notons simplement qu'à plusieurs reprises, les méthodes de Milligan et Cooper retrouvent deux classes dans les données. Cette classification en deux classes peut correspondre à d'une part des stations balnéaires ou à proximité d'une mer et à basse altitude, et d'autre part, aux stations plus enclavées et à haute altitude.

SCLUST	M1	M3	M4
k=7	20.54067	0.02262	0.93729
k=6	21.97853	0.02994	0.91742
k=5	23.29878	0.04691	0.86976
k=4	24.09909	0.05038	0.87722
k=3	26.63665	0.08011	0.80540
k=2	**33.59992**	0.08931	0.83494
k=1	-	-	-

TAB. 10.101 – Valeurs des indices de Milligan et Cooper pour SCLUST

10.5 Conclusion

Dans ce cas bien précis des variables de type intervalle, il semble que la méthode basée sur la modélisation milieu-longueur fournit de meilleurs résultats sur les données artificielles.

Nous n'entendons pas par là que la méthode utilisant le module DISS et que la "méthode de Hausdorff" ne fournissent pas des résultats convaincants. Néanmoins, les méthodes de détermination basées sur les tests d'hypothèse possèdent des indices peu significatifs et même parfois des indices qui ne le sont pas du tout.

Pour certains modèles comme celui des classes emboitées, la modélisation milieu-longueur est très intéressante.

Notons que les méthodes que nous avons mises en oeuvre sont très intéressantes car elle travaille directement sur les données symboliques. Le module DISS est intéressant pour les distances qu'il utilise car elles permettent la combinaison des variables de type intervalle avec d'autres types de variables symboliques.

Chapitre 11

Les variables multivaluées

11.1 Introduction

Le but de ce chapitre est de comparer les résultats obtenus à l'aide du module DISS avec ceux obtenus par S. Collès [2] sur les jeux de données étudiés par celle-ci.

Collès a étudié deux jeux de données. Le premier est un jeu qu'elle a créé et le deuxième décrit des boucles mérovingiennes.

Nous utiliserons le premier jeu pour comparer les différentes distances implémentées dans DISS pour les variables multivaluées. Pour rappel, ce sont les mêmes que celles utilisées pour les variables de type intervalle. Nous choisirons à nouveau la distance fournissant les meilleurs résultats s'il y en a une.

Nous rappelerons les résultats obtenus par S. Collès. Celle-ci avait utilisé les trois distances que nous avons rappelées à la fin du chapitre 3 à savoir :

- la distance de De Carvalho,
- la distance L1 et
- la distance L2.

Enfin, nous comparerons les résultats de détermination du nombre de classes obtenus pour les deux programmes.

11.2 Comparaison des distances implémentées dans DISS pour le jeu de données "ANIMAUX"

11.2.1 Informations sur le jeu de données

Ces données sont artificielles. Collès a pris 14 animaux et les a décrits de façon à distinguer à quelle espèce animale ceux-ci appartiennent.

Le jeu de données de départ contient donc 14 animaux dont

- 5 mammifères (chat, cheval, chien, ours, vache),

- 4 poissons (requin, saumon, thon, truite) et
- 5 oiseaux (aigle, hirondelle, oie, pigeon, poule)

décrits par 9 variables multivaluées. Les données forment donc trois classes.

"La première variable donne le mode de vie de ces animaux, c'est-à-dire s'ils sont sauvages, domestiques ou d'élevage. La deuxième fournit les modes de déplacement possibles et comprend quatre modalités : ailes, pattes, palmes et nageoires. La troisième variable décrit pour chaque animal sa race. Par exemple, la race du chien est le labrador. Cette variable contient 14 modalités. Il est évident que cette variable ne nous aidera pas pour notre problème de classification puisque chaque animal a sa propre race. Les quatrième et cinquième variables donnent respectivement le mode de reproduction (ovipare ou vivipare) et de respiration (poumons ou branchies). La sixième variable décrit le type de peau de chaque animal (poils, plumes ou écailles). La variable suivante dit dans quels milieux vit principalement l'animal (eau, terre, air). La huitième variable décrit le type d'alimentation (carnivore, granivore, herbivore, insectivore ou omnivore) et la dernière indique si ces animaux allaitent leurs petits, s'ils couvent leurs oeufs ou s'ils ne s'en occupent pas."

11.2.2 Comparaison

Nous avons premièrement jeté un rapide coup d'oeil sur les résultats obtenus par notre programme et nous avons remarqué que toutes les distances excepté une, la distance SO-5, fournissent des résultats similaires qui sont les suivants.

- Pour les méthodes de classification hiérarchiques :

		M1	M2	M3	M4	M5
saut minimum	+	3	3	3	3	3
saut maximum	+	3	3	3	3	3
Centroïde	+	3	3	3	3	3
Ward	+	3	3	3	3	3

TAB. 11.1 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques pour U-1, U-2, U-3, U-4, SO-1, SO-2, SO-3 et SO-4

- Pour la méthode de classification SCLUST :

		M1	M3	M4
SCLUST	+	3	3	3

TAB. 11.2 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par SCLUST pour U-1, U-2, U-3, U-4, SO-1, SO-2, SO-3 et SO-4

Nous remarquons que toutes les méthodes de classification qu'elles soient hiérarchiques ou pas retrouvent la bonne classification en trois classes. Les méthodes de détermination du nombre de classes sont également toutes unanimes et retrouvent le bon nombre de classes. Nous allons juste regarder s'il n'y a pas une méthode qui renvoie des indices plus significatifs que les autres.

En ce qui concerne la méthode SO-5, les résultats obtenus pour la détermination du nombre de classes sont les suivants.

- Pour les méthodes de classification hiérarchiques :

		M1	M2	M3	M4	M5
saut minimum	+	2	×	3	3	×
saut maximum	+	2	×	3	3	×
Centroïde	+	2	×	3	3	×
Ward	+	2	×	3	3	×

TAB. 11.3 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques pour SO-5

- Pour la méthodes de classification SCLUST :

		M1	M3	M4
SCLUST	+	2	3	3

TAB. 11.4 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par SCLUST pour SO-5

On constate que la classification est bien retrouvée pour toutes les méthodes de classification mais par contre les méthodes de Milligan et Cooper rencontrent des problèmes pour la détermination du nombre de classes dans les données.

11.2.3 La distance U-1

Les valeurs des indices sur les hiérarchies de partitions produites par la méthode du saut minimum sont reprises dans le tableau 11.5.

Les indices des méthodes de Milligan et Cooper associés aux partitions générées par SCLUST sont fournis par le tableau 11.6.

Nous constatons que tous les indices sont significatifs et ne permettent aucun doute pour les méthodes de Calinski et Harabaz (M1), du C-index (M3) et Gamma (M4). En ce qui concerne les méthodes de Duda et Hart (M2) et de Beale (M5), les indices n'atteignent pas les valeurs de rejet classiques.

saut minimum	M1	M2	M3	M4	M5
k=8	5.81730	2.92250	0.00000	1.00000	0.00000
k=7	6.07806	1.92856	0.01250	0.99297	1.00021
k=6	6.14077	1.43330	0.01739	0.98058	1.00833
k=5	7.48356	1.65343	0.00000	1.00000	0.75015
k=4	8.31366	2.04003	0.00478	0.99209	1.50041
k=3	**10.57951**	**1.38614**	**0.00000**	**1.00000**	**0.96738**
k=2	6.45410	3.14397	0.06496	0.76592	3.43807
k=1	-	2.32152	-	-	2.03929

TAB. 11.5 – Valeurs des indices de Milligan et Cooper pour la méthode du saut minimum : distance U-1

SCLUST	M1	M3	M4
k=5	6.65150	0.02432	0.98674
k=4	8.31366	0.00478	0.99230
k=3	**10.57951**	**0.00000**	**1.00000**
k=2	6.45410	0.06496	0.76592
k=1	-	-	-

TAB. 11.6 – Valeurs des indices de Milligan et Cooper pour SCLUST : distance U-1

11.2.4 La distance U-2

Les valeurs des indices sur les hiérarchies de partitions produites par la méthode du saut maximum sont reprises dans le tableau 11.7.

saut maximum	M1	M2	M3	M4	M5
k=8	2.60919	1.78766	0.02180	0.98107	0.86255
7	2.83263	1.65343	0.01786	0.98974	0.75015
6	3.06924	1.67667	0.01409	0.99800	1.08117
5	3.23095	1.61069	0.03239	0.96613	1.01801
4	3.71293	1.06674	0.04440	0.94575	0.71863
3	**4.78386**	**1.17044**	**0.00000**	**1.00000**	**0.79434**
2	4.34868	1.76108	0.01315	0.93045	1.51468
1	-	1.62472	-	-	1.37404

TAB. 11.7 – Valeurs des indices de Milligan et Cooper pour la méthode du saut minimum : distance U-2

Les indices des méthodes de Milligan et Cooper associés aux partitions générées par

SCLUST sont fournis par le tableau suivant :

SCLUST	M1	M3	M4
k=5	3.16156	0.02456	0.98674
k=4	3.80166	0.00210	0.99230
k=3	**4.78386**	**0.00000**	**1.00000**
k=2	4.34868	0.01315	0.93045
k=1	-	-	-

TAB. 11.8 – Valeurs des indices de Milligan et Cooper pour SCLUST : distance U-2

Comme pour la distance U-1, seuls les indices de Duda et Hart et ceux de Beale n'atteignent à nouveau pas les valeurs de rejet classiques.

11.2.5 Les distances U-3, U-4, SO-1 et S0-2

Les résultats fournis par ces distances sont sensiblement les mêmes que ceux obtenus par U-2. En effet, tous les indices sont significatifs excepté ceux des méthodes de Duda et Hart et ceux de Beale.

Remarque : En ce qui concerne les indices du C-index et de Gamma, la distance U-4 hésite entre trois ou quatre classes. En effet, les valeurs des indices obtenus sont exactement les mêmes et sont les valeurs idéales. Néanmoins, nous privilégions trois classes.

11.2.6 La distance SO-3

Les valeurs des indices sur les hiérarchies de partitions produites par la méthode du centroïde sont reprises dans le tableau 11.9.

Les indices des méthodes de Milligan et Cooper associés aux partitions générées par SCLUST sont fournis par le tableau 11.10.

Les indices sont cette fois-ci beaucoup plus significatifs. Pour la méthode de Calinski et Harabaz, le maximum est beaucoup plus marqué. Pour les méthodes basées sur les test d'hypothèse (M2) et (M5), les valeurs de rejet sont cette fois atteintes.

11.2.7 La distance SO-4

Cette distance fournit des résultats semblables à ceux obtenus pour la distance SO-3.

Centroïde	M1	M2	M3	M4	M5
k=8	123.19176	2.24601	0.00078	0.98999	1.41696
k=7	124.87207	2.92250	0.00071	0.99077	0.00000
k=6	136.29406	1.65343	0.00060	0.99202	0.75015
k=5	143.02192	1.93505	0.00000	1.00000	1.53308
k=4	112.46312	2.97279	0.00185	0.99209	3.77882
k=3	**125.19072**	**1.76162**	**0.00000**	**1.00000**	**1.33087**
k=2	18.82012	5.81285	0.10068	0.80490	27.97922
k=1	-	4.49247	-	-	5.94656

TAB. 11.9 – Valeurs des indices de Milligan et Cooper pour la méthode du saut minimum : distance U-2

SCLUST	M1	M3	M4
k=5	71.50343	0.00539	0.98678
k=4	121.81526	0.00072	0.99233
k=3	**125.19072**	**0.00000**	**1.00000**
k=2	18.82012	0.10068	0.80490
k=1	-	-	-

TAB. 11.10 – Valeurs des indices de Milligan et Cooper pour SCLUST : distance U-2

11.2.8 Conclusion

Les résultats obtenus par la méthode que nous avons mise en oeuvre sont très convaincants. Huit distances sur les neufs permettent de retrouver la bonne structure des données. Si nous devions choisir une distance sur les huit, ce serait soit SO-3 ou SO-4. En effet, ces distances donnent des indices beaucoup plus significatifs.

11.3 Comparaison de notre méthode avec celle de Séverine Collès

La méthode que nous avons mise en oeuvre et celle mise en oeuvre par S. COllès sont les mêmes. Seules les distances diffèrent. Il s'agit donc ici de comparer les distances utilisées.

11.3.1 "Données Animaux"

Résultats obtenus par Séverine Collès

◇ Distance de De Carvalho

Commençons par appliquer les méthodes de détermination du nombre de classes de Milligan et Cooper aux partitions générées par Sclust. Les résultats obtenus sont présentés dans le tableau suivant.

		M1	M3	M4
SCLUST	+	3	3	3

TAB. 11.11 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux partitions générées par Sclust : distance de De Carvalho

Sclust retrouve la bonne partition des données en trois classes et les trois méthodes de Milligan et Cooper fonctionnent bien. En effet, les indices des différentes méthodes prennent les valeurs suivantes.

Sclust	M1	M3	M4
8	4.65561	0.02011	0.98259
7	2.95738	0.03763	0.98704
6	4.49806	0.06616	0.96549
5	4.69607	0.08213	0.93414
4	7.56759	0.02703	0.95179
3	**10.51819**	**0.00000**	**0.95869**
2	4.64353	0.20468	0.68847
1	-	-	-

TAB. 11.12 – Valeurs des indices de Milligan et Cooper pour Sclust : distance de De Carvalho

Les méthodes de détermiation du nombre de classes appliquées aux partitions obtenues par les méthodes de classification hiérarchiques donnent les résultats présentés dans le tableau 11.13.

		M1	M2	M3	M4	M5
saut minimum	+	3	3	3,4	3,4	3
saut maximum	+	3	3	3,4	3,4	3
Centroïde	+	3	3	3,4	3,4	3
Ward	+	3	3	3,4	3,4	3

TAB. 11.13 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques : distance de De Carvalho

Les valeurs des différents indices des méthodes de Milligan et Cooper sont présentés dans la table 11.14 pour la méthode du saut minimum.

saut minimum	M1	M2	M3	M4	M5
k=8	5.44315	2.92250	0.00000	1.00000	0.00000
k=7	5.67194	1.92856	0.01250	0.98593	1.00021
k=6	5.70685	1.43330	0.02418	0.95146	1.00833
k=5	6.93793	1.65343	0.00606	0.97357	0.75015
k=4	8.34712	1.63612	**0.00000**	**1.00000**	1.04195
k=3	**10.51819**	1.56658	**0.00000**	**1.00000**	1.13090
k=2	7.15497	**3.00671**	0.03259	0.79510	**3.17976**
k=1	-	2.51950	-	-	2.26074

TAB. 11.14 – Valeurs des indices de Milligan et Cooper pour la méthode du saut minimum : distance de De Carvalho

Les méthodes de détermination du nombre de classes retrouvent toutes la présence de trois classes dans les données. Notons cependant que les méthodes du C-index (M3) et Gamma (M4) atteignent leurs valeurs idéales pour trois et quatre classes. La classe supplémentaire quand nous passons de trois à quatre classes ne contient qu'un seul individu (la poule).

Le fait que ces deux méthodes atteignent respectivement les valeurs 0 et 1 indique que pour chacune des partitions, aucun objet n'est mal placé, dans le sens où un objet est mal classé si sa distance à la classe à laquelle il appartient est plus grande que sa distance à une autre classe. Ce qui peut se comprendre intuitivement ici en remarquant que la poule est l'oiseau le plus différent des autres puisqu'il vit au sol et qu'il est d'élevage.

Les résultats obtenus pour les méthodes du saut maximum, du centroïde et de Ward sont fort semblables à ceux obtenus pour la méthode du lien simple. Toutes les méthodes forment bien les classes recherchées. L'indice de Calinski et Harabasz (M1) atteint un maximum en trois classes. Les indices de Duda et Hart (M2) et de Beale (M5) n'atteignent

pas les valeurs qui permettent de rejeter l'hypothèse de fusion des classes mais atteignent un maximum en $k = 2$. Les méthodes du C-index (M3) et Gamma (M4) atteignent pour toutes les méthodes hiérarchiques leurs valeurs idéales, c'est-à-dire respectivement 0 et 1, pour trois et quatre classes.

◇ **Distance L_2**

Les quatre méthodes hiérarchiques nous donnent les résultats présentés dans le tableau 11.15 pour les méthodes de détermination du nombre de classes.

		M1	M2	M3	M4	M5
saut minimum	+	3	3	3,4	3,4	3
saut maximum	+	3	3	3,4	3,4	3
Centroïde	+	3	3	3,4	3,4	3
Ward	+	3	3	3,4	3,4	3

TAB. 11.15 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques

Regardons maintenant les valeurs prises par les indices des cinq méthodes de détermination du nombre de classes de Milligan et Cooper appliquées à la hiérarchie de partitions générée par la méthode du centroïde. Les résultats sont repris dans le tableau suivant.

Centroïde	M1	M2	M3	M4	M5
k= 8	5.34694	2.92250	0.00000	1.00000	0.00000
k=7	4 5.92667	1.65343	0.00000	1.00000	0.75015
k=6	6.41758	1.92856	0.01075	0.99199	1.00021
k=5	7.29419	1.49740	0.00913	0.99325	0.91692
k=4	8.36798	1.43330	**0.00000**	**1.00000**	1.00833
k=3	**10.51533**	1.62399	**0.00000**	**1.00000**	1.18706
k=2	7.03967	**3.03494**	0.05045	0.79774	**3.23123**
k=1	-	2.48793	-	-	2.22431

TAB. 11.16 – Valeurs des indices de Milligan et Cooper pour la méthode du centroïde

Toutes les méthodes mettent en évidence une classification en trois classes. A nouveau, les méthodes du C-index (M3) et Gamma (M4) indiquent la présence de trois ou quatre classes.

Les méthodes de détermination du nombre de classes de Milligan et Cooper appliquées aux partitions générées par Sclust donnent les résultats repris dans le tableau 11.17.

		M1	M3	M4
SCLUST	+	3	3,4	3

TAB. 11.17 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux partitions générées par Sclust

Pour commencer, nous constatons que Sclust retrouve bien les trois classes naturelles. Seule la méthode du C-index (M3) hésite entre trois et quatre classes. Les deux autres méthodes détectent trois classes dans les données.

SCLUST	M1	M3	M4
8	5.34694	0.00000	1.00000
7	3.45942	0.03763	0.96198
6	6.08000	0.02151	0.96607
5	6.79762	0.01210	0.96888
4	8.36798	**0.00000**	0.99872
3	**10.51533**	**0.00000**	**1.00000**
2	6.58696	0.09384	0.67246
1	-	-	-

TAB. 11.18 – Valeurs des indices de Milligan et Cooper pour Sclust

En conclusion, pour les partitions générées avec SCLUST, les résultats obtenus pour la distance L_2 sont les mêmes que pour la distance de De Carvalho. A noter cependant que la méthode du C-index (M3) appliquée aux partitions générées par Sclust atteint sa valeur idéale pour trois et quatre classes. Une nouvelle fois, la poule forme la quatrième classe.

◇ **Distance L_1**

Les résultats obtenus pour les méthodes de détermination du nombre de classes associées aux méthodes de classification hiérarchiques étaient les suivants.

		M1	M2	M3	M4	M5
saut minimum	+	3	3	3,5	3,5	3
saut maximum	+	3	3	3,4,5	3,4,5	3
Centroïde	+	3	3	3,4,5	3,4,5	3
Ward	+	3	3	3,4	3,4	3

TAB. 11.19 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques

Alors que les méthodes de Calinski et Harabasz (M1), de Duda et Hart (M2) et de Beale (M5) s'accordent pour dire qu'il y a trois classes dans les données, les méthodes du C-index (M3) et Gamma (M4) hésitent entre trois et quatre classes pour les méthodes du saut minimum et de Ward et entre trois, quatre et cinq classes pour les méthodes du saut maximum et du centroïde.

Les différentes valeurs obtenues pour les méthodes de détermination du nombre de classes pour les partitions produites par la méthode du saut maximum sont reprises dans le tableau 11.20.

saut maximum	M1	M2	M3	M4	M5
8	5.52770	2.92250	0.00000	1.00000	0.00000
7	6.13333	1.65343	0.00000	1.00000	0.75015
6	5.98442	1.90922	0.01942	0.98545	1.33370
5	7.06122	0.79579	**0.00000**	**1.00000**	0.54018
4	8.30570	1.76096	**0.00000**	**1.00000**	1.16699
3	**10.49177**	1.51299	**0.00000**	**1.00000**	1.08036
2	6.84682	**3.05755**	0.04776	0.79510	**3.27305**
1	-	2.43427	-	-	2.16337

TAB. 11.20 – Valeurs des indices de Milligan et Cooper pour la méthode du saut maximum

Regardons les résultats obtenus en appliquant les méthodes de détermination du nombre de classes aux partitions générées par Sclust dans le tableau 11.21.

		M1	M3	M4
SCLUST	+	3	3,4,5	3,4,5

TAB. 11.21 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux partitions générées par Sclust

Les méthodes du C-index (M3) et Gamma (M4) indiquent la présence de trois, quatre ou cinq classes dans les données. Comme pour la distance L_2, la classe supplémentaire en passant de trois à quatre classes est formée de la poule. En ce qui concerne la partition en cinq classes, elle n'est pas du tout intéressante puisque la classe supplémentaire formée par Sclust ne contient aucun élément.

Voici maintenant la valeur des différents indices des méthodes de détermination du nombre de classes.

SCLUST	M1	M3	M4
8	3.20594	0.05000	0.94231
7	5.42972	0.05000	0.94169
6	5.90857	0.05102	0.98415
5	7.06122	**0.00000**	**1.00000**
4	8.30570	**0.00000**	**1.00000**
3	**10.49177**	**0.00000**	**1.00000**
2	6.45967	0.09145	0.66218
1	-	-	-

TAB. 11.22 – Valeurs des indices de Milligan et Cooper pour Sclust

◇ Conclusion

Signalons tout d'abord que toutes les méthodes de classification retrouvent la bonne partition des données en trois classes pour les trois distances.

Alors que les méthodes de Calinski et Harabasz (M1), de Duda et Hart (M2) et de Beale (M5) indiquent clairement la présence de trois classes, les méthodes du C-index (M3) et Gamma (M4) hésitent entre trois, quatre et cinq classes.

Notons que, bien que la classification en trois classes nous semble la plus normale, les classifications en quatre et cinq classes peuvent aussi être intéressantes.

Résultats obtenus avec DISS

Les résultats obtenus ont été présentés dans la section précédente.

Conclusion

Aucune des deux méthodes n'est meilleure que l'autre. Il est une fois de plus très difficile de tirer des conclusions.

Constatons toutefois que dans le module DISS, la distance SO-5 doit être bannie. Toutes les autres distances fournissent des résultats similaires.

Remarquons néanmoins que les indices obtenus pour les méthodes basées sur les tests d'hypothèse n'atteignent jamais les valeurs de rejet pour les distances utilisées par Séverine Collès ce qui n'est pas le cas lorsque nous utilisons les distances SO-3 et SO-4 dans le module DISS.

Nous constatons également que les méthodes du C-index (M3) et Gamma (M4) indiquent la présence de trois, quatre ou cinq classes pour toutes les distances étudiées par Séverine Collès. La partition en quatre classes isole, bien souvent, dans une quatrième classe la poule ou l'oie. Pour obtenir la partition en cinq classes, l'oie ou la poule est isolée des autres oiseaux.

11.3.2 Boucles mérovingiennes datant du 6-8ème siècle après Jésus-Christ [2]

Nous allons pour le module DISS uniquement reprendre les résultats obtenus pour la distance SO-4 vu que cette distance fournit de bien meilleurs résultats pour le jeu de données précédent. Nous regarderons également les résultats pour la distance U-2 car il s'agit de celle utilisée pour les variables de type intervalle. Les autres distances fournissent de toute façon des résultats semblables.

Informations sur le jeu de données

Ce jeu de données est téléchargeable à l'adresse

`http://www-rocq.inria.fr/sodas/WP6/data/data.html`

Il contient 59 boucles mérovingiennes décrites par six variables multivaluées avec en tout 23 catégories (voir tableau 11.23).

Variables	Catégories
Fixation	clous de fer, bossettes de bronze, aucune
Bordures	bords mouvementés, répétition de motifs, frises géométriques
Damasquinure	bichrome, placage prédominant, incrustation dominante, monochrome argent
Fond	plaque d'argent, hachures, trame géométrique
Incrustation	filiforme, bande hachurée, bande pointillée, ruban plein
Plaque	arabesque, grande taille, dorsale carrée, motifs animaliers, tresse, ronde

TAB. 11.23 – Variables symboliques sur les boucles

Les archéologues ont retrouvé deux classes (A et B) dans ce jeu de données.

Une partition de ces données a été obtenue par Leredde en 1979 en utilisant une classification hiérarchique basée sur les connaissances des archéologues. Représentons chaque

Classes	Individus présents dans la classe
A	3, 4, 5, 7, 9, 11, 14, 15, 17, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 31, 32, 33, 34, 35, 36, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59
B	1, 2, 6, 8, 10, 12, 13, 16, 18, 19, 29, 37, 38, 39, 40, 41, 42, 43, 44, 45

TAB. 11.24 – Description des deux classes obtenues par les archéologues

boucle par un numéro. Les deux classes sont celles données dans le tableau 11.24.

La première classe contient des individus dont la variable "damasquinure" prend les modalités "incrustation dominante" et "monochrome argent" et dont la variable "fond" prend principalement la modalité "hachures". Les individus de la seconde classe par contre prennent pour la première de ces deux variables les modalités "bichrome" et "placage prédominant" et pour la seconde variable "plaque d'argent".[11]
Un exemple de boucles mérovingiennes de chaque classe est fournit à la figure 11.1.

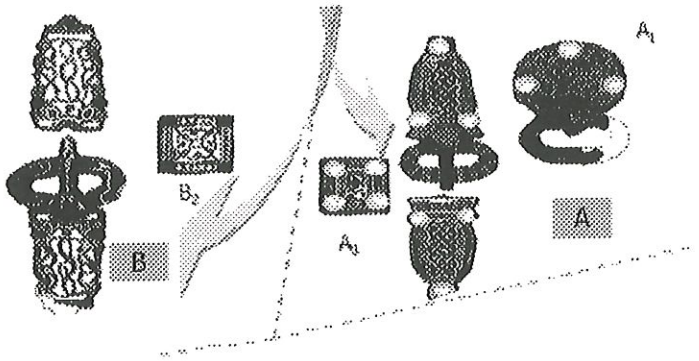


FIG. 11.1 – Exemple de boucles mérovingiennes

Remarquons que les archéologues ont aussi défini une classification de ces individus en sept classes. Cette classification correspond en fait à une subdivision des deux classes ci-dessus. Elle est présentée dans le tableau 11.25.

Classes	Individus présents dans la classe
A_1	15, 23, 47, 55, 56, 57, 58
A_2	5
A_3	3, 4, 9, 14, 20, 21, 22, 24, 25, 26, 28, 30, 31, 32, 33, 35, 36, 46, 49, 53
A_4	7, 11, 17, 27, 34, 48, 50, 51, 52, 59
B_1	8, 45
B_2	2, 6, 12, 13, 18, 19, 29, 44
B_3	1, 10, 16, 37, 38, 39, 40, 41, 42, 43

TAB. 11.25 – Description des sept classes obtenues par les archéologues

Résultats obtenus par Séverine Collès

◇ Distance de De Carvalho

Considérons la distance de De Carvalho et appliquons les méthodes de détermination du nombre de classes de Milligan et Cooper sur les partitions générées par les quatres méthodes de classification hiérarchiques. Les résultats obtenus sont repris dans le tableau 11.26.

		M1	M2	M3	M4	M5
saut minimum	−	2	2	2	2	2
saut maximum	+	2	2	2	2	2,6
Centroïde	−	2	2	2	2	2
Ward	+	2	2	2	2	2,4

TAB. 11.26 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques

Les méthodes du saut minimum et du centroïde ne retrouvent pas la classification naturelle des boucles puisqu'elles placent la boucle 20 dans le mauvais groupe. Pour ces deux méthodes de classification, les méthodes de détermination du nombre de classes ne sont donc pas appliquées à la bonne partition. On constate que toutes les méthodes de Milligan et Cooper préconisent le bon nombre de classes dans les données.

Les valeurs des indices des différentes méthodes de Milligan et Cooper sont présentées dans le tableau 11.27 en ce qui concerne la méthode du saut maximum.

Tous les indices sont significatifs et ne laissent aucun doute quant à la présence de deux classes dans les données excepté l'indice de Beale (M5) pour lequel aucune des deux valeurs de rejet habituellement utilisées (4.61 ou 5.30) n'est atteinte. Cependant, il atteint deux valeurs fort proches et assez importantes en $k = 5$ et $k = 1$, ce qui nous laisse présumer la

Saut maximum	M1	M2	M3	M4	M5
8	31.07487	2.39966	0.00790	0.99238	3.72846
7	33.83576	2.35439	0.00860	0.99117	0.00000
6	34.66279	2.28551	0.01338	0.98565	2.34051
5	28.87829	3.56668	0.03270	0.94615	**3.59244**
4	34.53483	1.48947	0.03986	0.93230	1.20779
3	29.45371	2.91271	0.02761	0.95592	1.88482
2	**51.47556**	**1.53622**	**0.00932**	**0.99010**	**1.24133**
1	-	5.27061	-	-	3.20213

TAB. 11.27 – Valeurs des indices de Milligan et Cooper pour la méthode du saut maximum

présence de deux ou six classes dans les données. La classification en six classes est pertinente. Pour rappel, les archéologues avait également réalisé une partition en sept classes.

Appliquons maintenant les méthodes de détermination du nombre de classes de Milligan et Cooper sur les hiérarchies de partitions générées par SCLUST.

		M1	M3	M4
SCLUST	–	2	2	2

TAB. 11.28 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à Sclust

Sclust ne retrouve pas la partition en deux classes décrite par les archéologues. Cependant, il n'y a qu'un seul individu mal classé. Il s'agit de la boucle 20 qui semble avoir plusieurs caractéristiques communes à la fois avec les individus de la classe A et les individus de la classe B. La partition fournie par Sclust n'est pas pour autant mauvaise.

Les indices des méthodes de Milligan et Cooper prennent les valeurs suivantes :

SCLUST	M1	M3	M4
8	28.52599	0.03049	0.95382
7	30.37194	0.04221	0.95325
6	23.36924	0.05127	0.92168
5	31.03050	0.04869	0.93004
4	32.66536	0.05707	0.95822
3	41.24019	0.03788	0.96124
2	**51.13007**	**0.01201**	**0.99917**
1	-	-	-

TAB. 11.29 – Valeurs des indices de Milligan et Cooper pour Sclust

La méthode de Calinski et Harabasz (M1) atteint un maximum absolu en deux classes. De même, les méthodes du C-index (M3) et Gamma (M4) présentent respectivement un minimum et un maximum pour ce même nombre de classes.

◇ Distance L2

Lorsque nous appliquons les méthodes de détermination du nombre de classes de Milligan et Cooper aux partitions générées par les quatre méthodes de classification hiérarchiques, nous obtenons les résultats repris dans le tableau 11.30.

		M1	M2	M3	M4	M5
saut minimum	–	2	2	2	2	2
saut maximum	–	2	2	2	2	6
Centroïde	–	2	2	2	2	2
Ward	–	2	2	5	5	4

TAB. 11.30 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques

Aucune des méthodes classification ne retrouvent la partition des données formulée par les archéologues. Pour les méthodes du saut minimum et du centroïde, toutes les méthodes s'accordent pour trouver deux classes dans les données malgré le fait que l'individu 20 ne soit pas placé dans la même classe que celle où les archéologues l'ont placé. Quant à la méthode de Ward, les résultats obtenus sont moins clairs, ce qui peut être dû au fait que non seulement l'individu 20 est mal classé mais également l'individu 35.

Pour la méthode du saut maximum, quelques différences sont à noter par rapport à la distance de De Carvalho.

Premièrement, la méthode ne retrouve plus la partition en deux classes décrite par les archéologues puisque l'individu 20 est placé dans la classe B au lieu de la classe A.

Deuxièmement, alors que les méthodes de Calinski et Harabasz (M1), de Duda et Hart (M2), du C-index (M3) et Gamma (M4) retrouvent toujours bien deux classes, la méthode de Beale (M5) aurait plutôt tendance à donner six classes.

Au niveau des six classes, l'une reprend les groupes $A1$ et $A2$, une autre correspond au groupe $A3$ sans l'individu 20 qui forme une classe à lui seul, la quatrième classe est le groupe $A4$ et les deux dernières divisent le groupe B mais nous ne pouvons distinguer aucun rapport avec la classification faite par les archéologues.

Les valeurs des indices trouvées pour la méthode du saut minimum sont données dans le tableau 11.31.

Appliquons maintenant les méthodes de détermination du nombre de classes de Milligan et Cooper aux partitions générées par Sclust.

Saut maximum	M1	M2	M3	M4	M5
8	29.96255	2.39966	0.00744	0.99387	3.72846
7	31.65414	0.81396	0.01461	0.98313	0.78073
6	34.63818	2.08800	0.01643	0.98027	2.24749
5	29.01056	4.40728	0.03381	0.94601	**6.59618**
4	31.86908	1.13986	0.05267	0.90653	0.97276
3	26.71250	2.99006	0.03080	0.94924	1.96205
2	**46.39720**	1.32172	**0.01376**	**0.98480**	1.08243
1	-	**4.90141**	-	-	2.88622

TAB. 11.31 – Valeurs des indices de Milligan et Cooper pour la méthode du saut maximum

Sclust ne retrouve pas la bonne partition en deux classes mais seul l'individu 20 se trouve dans un mauvais groupe. Les méthodes de Calinski et Harabasz (M1), du C-index (M3) et Gamma (M4) retrouvent toutes deux classes dans les données.

		M1	M3	M4
SCLUST	–	2	2	2

TAB. 11.32 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à Sclust

En effet, les indices des différentes méthodes prennent les valeurs reprises dans le tableau 11.33.

SCLUST	M1	M3	M4
6	34.27613	0.05379	0.92349
5	24.55655	0.06042	0.93119
4	39.80523	0.03925	0.96119
3	41.26604	0.03353	0.94161
2	**46.39720**	**0.01376**	**0.98784**
1	-	-	-

TAB. 11.33 – Valeurs des indices de Milligan et Cooper pour Sclust

◇ Distance L1

Les méthodes de détermination du nombre de classes appliquées aux partitions générées par les quatre méthodes de classification hiérarchiques fournissent les résultats suivants :

		M1	M2	M3	M4	M5
saut minimum	–	2	2	2	2	2
saut maximum	–	2	2	2	2	6
Centroïde	–	2	2	2	2	2
Ward	–	2	2	2	2	4

TAB. 11.34 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques

Aucune des méthodes ne retrouve ici la classification donnée par les archéologues. A chaque fois, c'est l'individu 20 qui pose problème. Toutes les méthodes indiquent la présence de deux classes naturelles dans les données pour les méthodes du saut minimum et du centroïde. Pour la méthode du saut maximum, l'indice de Beale (M5) retrouve six classes.

Notons aussi que la méthode du centroïde place les éléments 20 et 35 dans le mauvais groupe en ce qui concerne la classification en deux classes.

Les valeurs des indices des méthodes de détermination du nombre de classes sont données pour la méthode de Ward dans le tableau suivant :

Ward	M1	M2	M3	M4	M5
8	28.02594	1.46563	0.02551	0.96836	1.32603
7	28.77218	1.69869	0.03059	0.95963	1.56108
6	30.63448	1.98742	0.03620	0.94777	2.05914
5	33.30967	1.39366	0.04886	0.92312	1.14068
4	36.56043	1.70319	0.04055	0.92956	1.36556
3	39.29677	3.44566	0.04663	0.92363	**3.38329**
2	**47.02353**	2.58917	**0.00771**	**0.99482**	1.66575
1	-	**4.94889**	-	-	2.92518

TAB. 11.35 – Valeurs des indices de Milligan et Cooper pour la méthode de Ward

Pour cette méthode, l'indice de Beale (M5) met en évidence une classification en quatre classes. Cette partition est la suivante. La première classe reprend les groupes $A1$ et $A2$. La deuxième correspond au groupe $A3$. La troisième classe est le groupe $A4$ et la dernière classe correspond au groupe B.

Voici les résultats obtenus par les méthodes de Milligan et Cooper appliquées à Sclust.

		M1	M3	M4
SCLUST	–	2	2	2

TAB. 11.36 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à Sclust

La partition des archéologues n'est pas retrouvée, seul l'élément 20 est une nouvelle fois mal placé. Mais toutes les méthodes retrouvent le bon nombre de classes. En effet, les indices des méthodes prennent les valeurs suivantes :

SCLUST	M1	M3	M4
6	34.27613	0.05379	0.92349
5	24.55655	0.06042	0.93119
4	39.80523	0.03925	0.96119
3	41.26604	0.03353	0.94161
2	**46.39720**	**0.01376**	**0.98784**
1	-	-	-

TAB. 11.37 – Valeurs des indices de Milligan et Cooper pour Sclust

◇ Conclusion

La distance de De Carvalho est celle qui donne les meilleurs résultats pour cet exemple car seuls deux cas ne déterminent pas le bon nombre de classes dans les données. Il s'agit de la méthode de Beale (M5) sur les partitions générées par les méthodes de Ward et du saut maximum. Et c'est aussi la seule qui retrouve la partition en deux classes décrite par les archéologues.

La distance L_2 est celle qui fournit les moins bons résultats.

Notons aussi que la classification des boucles mérovingiennes par les archéologues n'est probablement pas basée sur les mêmes critères que ceux qu'utilisent Sclust et les quatre méthodes hiérarchiques. Ceci expliquerait pourquoi l'individu 20, bien que placé dans la classe A par les archéologues, se retrouve la plupart du temps dans la classe B.

Pour ce jeu de données sur les boucles mérovingiennes datant du 6-8ème siècle après Jésus-Christ, nous concluons à la présence de deux classes naturelles dans les données.

Résultats obtenus avec le module DISS

◇ La distance U-2

Les méthodes de détermination du nombre de classes appliquées aux partitions générées par les quatre méthodes de classification hiérarchiques fournissent les résultats repris dans le tableau 11.38.

		M1	M2	M3	M4	M5
saut minimum	+	2	2	2	2	2
saut maximum	+	2	2	2	2	2
Centroïde	+	2	2	2	2	2
Ward	+	2	2	2	2	2

TAB. 11.38 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques

Toutes les méthodes retrouvent la classification donnée par les archéologues lorsque deux classes sont requises. Toutes les méthodes indiquent la présence de deux classes naturelles dans les données quelque soit la méthode de classification.

Les valeurs des indices des méthodes de détermination du nombre de classes sont données pour la méthode du saut minimum dans le tableau suivant :

Saut minimum	M1	M2	M3	M4	M5
k=8	9.19026	1.37457	0.05717	0.88059	1.22324
k=7	8.34434	2.64441	0.04388	0.90664	2.38359
k=6	9.62610	-0.74347	0.03373	0.92752	0.13873
k=5	10.57134	-0.17808	0.02628	0.94847	0.33778
k=4	12.52565	0.57249	0.02078	0.96273	0.64219
k=3	17.80465	-0.76131	0.00907	0.98622	0.14413
k=2	**33.59689**	**-0.16203**	**0.00133**	**0.99919**	**0.28756**
k=1	-	3.78718	-	-	2.08995

TAB. 11.39 – Valeurs des indices de Milligan et Cooper pour la méthode de Ward

Tous ces indices sont significatifs. Notons que l'indice de Beale n'atteint pas la valeur de rejet classique.

Voici les résultats obtenus par les méthodes de Milligan et Cooper appliquées à Sclust.

		M1	M3	M4
SCLUST	–	2	2	2

TAB. 11.40 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à Sclust

La partition des archéologues n'est pas retrouvée, l'élément 20 est mal placé. Mais toutes les méthodes retrouvent le bon nombre de classes.

Les indices des méthodes de Milligan et Cooper prennent les valeurs suivantes.

SCLUST	M1	M3	M4
k=5	18.81105	0.07143	0.85709
k=4	22.92292	0.08427	0.82540
k=3	26.24067	0.07124	0.85723
k=2	31.76836	0.02083	0.97231
k=1	-	-	-

TAB. 11.41 – Valeurs des indices de Milligan et Cooper pour Sclust

◇ La distance SO-4

Les méthodes de détermination du nombre de classes appliquées aux partitions générées par les quatre méthodes de classification hiérarchiques fournissent les résultats suivants.

		M1	M2	M3	M4	M5
saut minimum	–	5	×	×	×	×
saut maximum	–	2	6,2	?	?	2
Centroïde	–	2	3,2	?	2	2
Ward	–	2	3,2	?	2	2

TAB. 11.42 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques

Aucune des méthodes ne retrouvent la classification réalisée par les archéologues lorsque deux classes sont demandées. Les résultats indiqués par les différentes méthodes ne sont pas ceux attendus. Pour la méthode du saut minimum, les indices sont très peu significatifs et feraient penser à un jeu de données sans structure. Pour les trois autres méthodes, on retrouve bien souvent la présence de deux classes dans les données. Seule le méthode du C-index (M3) ne permet aucune conclusion.

Les valeurs des indices des méthodes de détermination du nombre de classes sont données pour la méthode de Ward dans le tableau 11.43.

Tous ces indices sont significatifs. Notons que l'indice de Beale n'atteint pas la valeur de rejet.

Ward	M1	M2	M3	M4	M5
k=8	137.39049	2.23837	0.00983	0.88779	2.63185
k=7	125.21204	3.04109	0.01526	0.84947	2.77130
k=6	136.47253	2.98025	0.01571	0.84145	2.92907
k=5	148.50572	2.33420	0.01768	0.82140	2.50119
k=4	195.12620	-0.43775	0.01877	0.84462	0.19456
k=3	234.74259	**2.62983**	0.02422	0.82402	2.25887
k=2	**285.02895**	4.71952	0.02790	**0.93820**	**3.50584**
k=1	-	10.40358	-	-	17.73073

TAB. 11.43 – Valeurs des indices de Milligan et Cooper pour la méthode de Ward

Voici les résultats obtenus par les méthodes de Milligan et Cooper appliquées à Sclust :

		M1	M3	M4
SCLUST	–	2	2	2

TAB. 11.44 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à Sclust

La partition des archéologues n'est pas retrouvée, l'élément 20 est mal placé. Toutes les méthodes de Milligan et Cooper retrouvent néanmoins le bon nombre de classes.

Les valeurs des indices des méthodes de détermination du nombre de classes prennent les valeurs suivantes.

SCLUST	M1	M3	M4
k=5	167.77639	0.01583	0.76586
k=4	181.64869	0.02554	0.72717
k=3	267.37810	0.02002	0.81698
k=2	**353.74933**	**0.01108**	**0.95003**
k=1	-	-	-

TAB. 11.45 – Valeurs des indices de Milligan et Cooper pour Sclust

◇ Conclusion

La distance U-2 est la distance fournissant les meilleurs résultats. En effet, grâce à celle-ci les méthodes de classification hiérarchiques retrouvent la bonne partition des données. Les indices de détermination du nombre de classes sont également très significatifs.

Pour la méthode SCLUST, un seul élément est mal placé et cela ne doit pas avoir de trop grandes conséquences sur les indices de Milligan et Cooper.

Chapitre 12

Les variables modales

12.1 Introduction

Le but de ce chapitre est de comparer les résultats obtenus à l'aide du module DISS avec ceux obtenus par S. Collès [2] sur les jeux de données étudiés par celle-ci.

Collès a étudié deux jeux de données décrits par des variables modales. Les deux jeux sont des jeux de données réels.

Nous rappelerons les résultats obtenus par Séverine Collès. Comme pour les variables multivaluées, elle avait utilisé les trois distances que nous avons rappelées à la fin du chapitre 3 à savoir :

- la distance de De Carvalho,
- la distance L1 et
- la distance L2.

Nous étudierons ensuite les résultats que nous obtenons avec l'utilisation du module DISS. Notons que nous avons su étudier les résultats que pour deux des six distances étudiées dans le chapitre 6. Il s'agit des distances

- PU-1 associée à la distance de Minkowski et
- PU-2.

Après plusieurs tentatives pour les quatre autres distances, nous avons dû admettre que le logiciel SODAS bouclait pour ces distances. Nous n'étudierons donc que les résultats obtenus pour les deux distances citées ci-dessus.

12.2 Comparaison de nos distances avec celles utilisées par Séverine Collès

12.2.1 Magasins e-Fashion

Informations sur le jeu de données [2]

Le jeu de données regroupe sur trois années (1999, 2000 et 2001) les ventes d'une chaîne de 13 magasins de vêtements et accessoires, répartis dans six pays différents. Ce jeu de données est téléchargeable à l'adresse suivante :

`http://www.ceremade.dauphine.fr/~touati/exemples.htm`.

Les 13 individus sont les magasins de Paris 6ème, Lyon, Rome, Barcelone, Toulouse, Aix-Marseille, Madrid, Berlin, Milan, Bruxelles, Paris 15ème, Paris 8ème et Londres.

Les 8 différentes variables modales sélectionnées dans le jeu de données sont :

	Nom	Nombre de modalités
1	étiquette article	153
2	catégorie	31
3	famille du produit	12
4	étiquette couleur	160
5	gamme de couleur	17
6	mois	12
7	niveau de vente	5
8	numéro	5

Les variables "étiquette article", "famille du produit", "étiquette couleur", "gamme de couleur" et "catégorie" décrivent les articles vendus dans les magasins. La variable "étiquette article" peut prendre par exemple les modalités suivantes :

- Collier avec tiges longues,
- Echarpe en Viscose avec rectangles,
- Blouson Esprit Rivoli 5 poches,
- Veste Tobaggo Shetland,

et elle leur associe la proportion des ventes réalisées pour chaque article dans le magasin considéré.

La variable "famille produit" associe aussi la proportion des ventes mais en fonction des types de produits comme par exemple :

- Robes,
- Pulls,
- Sweats et T-shirts,
- Chemisiers...

Les variables "étiquette couleur", "gamme de couleur" et "catégories" indiquent aussi la proportion des ventes mais en fonction des couleurs des articles vendus (Bleu Marine Oxford,

Bleu Indigo, Gris Chine, ...), en fonction des gammes de couleurs (Bleu, Gris, Marron, Bordeaux, Rouge, ...) et en fonction des catégories des produits vendus (Bermudas ou Shorts, Pantalons Stricts, Habits - Soirée, Col Roulé, ...).
La variable "mois" indique la proportion des ventes effectuées durant chaque mois de l'année. La variable "numéro" peut prendre les valeurs 2, 3, 4, 5, 6 qui correspondent au numéro de la promotion qui a été effectuée (promotion dans le magasin, publicité à la radio, envoi de publicités par email, ...). Le niveau de vente reprend les possibilités suivantes : très faible, faible, nul, moyen et grand. Il indique donc les résultats de vente réalisés pour chaque magasin.

Notons aussi que nous ne connaissons rien sur une éventuelle classification naturelle des données.

Résultats obtenus par Séverine Collès

◇ Distance de De Carvalho

Appliquons les méthodes de détermination du nombre de classes de Milligan et Cooper sur les hiérarchies de partitions générées par les quatre méthodes de classification classiques.

Les résultats obtenus sont les suivants :

	M1	M2	M3	M4	M5
saut minimum	2	2	2	2	2
saut maximum	2	2	2	2	2
Centroïde	2	2	2	2	2
Ward	2	2	2	2	2

TAB. 12.1 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques

Toutes les méthodes de Milligan et Cooper indiquent la présence de deux classes dans les données. Toutes les méthodes de classification retrouvent la même partition en deux classes. L'une de ces classes ne contient que le magasin de Londres. Nous verrons plus tard pourquoi ce magasin se distingue de tous les autres.

Regardons les valeurs que prennent les différentes méthodes de Milligan et Cooper pour la méthode de Ward dans le tableau 12.2.

Les résultats obtenus par les méthodes de détermination du nombre de classes appliquées aux partitions produites par Sclust sont présentés dans le tableau 12.3.

Alors que la méthode Gamma (M4) n'indique aucune structure dans les données, la méthode de Calinski et Harabasz (M1) met en évidence deux classes et celle du C-index

Ward	M1	M2	M3	M4	M5
12	13.37300	2.74619	0.00000	1.00000	0.00000
11	9.16062	2.74619	0.00000	1.00000	0.00000
10	8.49961	2.74619	0.00000	1.00000	0.00000
9	8.68826	2.74619	0.00000	1.00000	0.00000
8	8.68813	2.74619	0.00192	0.95616	0.00000
7	8.72421	2.62949	0.00406	0.91147	2.88679
6	8.83188	1.57477	0.00699	0.90769	1.20339
5	9.21637	1.86857	0.01319	0.86114	1.39562
4	10.09056	1.45145	0.01311	0.87364	1.20420
3	11.77580	1.40491	0.03753	0.74332	1.18708
2	**11.82885**	2.20615	**0.00000**	**1.00000**	2.00135
1	-	**3.33614**	-	-	**3.61685**

TAB. 12.2 – Valeurs des indices de Milligan et Cooper pour la méthode de Ward

	M1	M3	M4
SCLUST	2	4	×

TAB. 12.3 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à Sclust

(M3) quatre classes.

En effet, les indices prennent les valeurs suivantes :

SCLUST	M1	M3	M4
6	1.79292	0.11571	0.25692
5	1.27726	0.24985	0.16712
4	2.79681	**0.09853**	0.34385
3	2.02706	0.33618	0.29278
2	**2.95234**	0.39883	0.35245
1	-	-	-

TAB. 12.4 – Valeurs des indices de Milligan et Cooper pour Sclust

Voici la partition en deux classes retrouvée par Sclust.

```
Classe :   1  Cardinal :      7
================================
 (0)  "e-Fashion Paris 6ème"    [0.3]  (5)  "e-Fashion Aix-Marseille" [0.4]
 (8)  "e-Fashion Milano"        [0.1]  (9)  "e-Fashion Bruxelles"  [0.3]
 (10) "e-Fashion Paris 15ème"   [0.3]  (11) "e-Fashion Paris 8ème" [0.2]
 (12) "e-Fashion London"        [5.4]

Classe :   2  Cardinal :      6
================================
 (1) "e-Fashion Lyon"           [0.5]  (2) "e-Fashion Roma"  [2.8]
 (3) "e-Fashion Barcelona"      [0.6]  (4) "e-Fashion Toulouse" [0.7]
 (6) "e-Fashion Madrid"         [0.7]  (7) "e-Fashion Berlin" [0.7]
```

Notons que le chiffre entre crochet indique la distance entre l'individu et son prototype. Remarquons que le magasin e-Fashion de Londres est fort éloigné du prototype de sa classe par rapport aux autres individus. Cette remarque est aussi valable pour l'e-Fashion de Rome.

Notons que si nous exécutons plusieurs fois le programme, les résultats obtenus sont à chaque fois différents puisque la partition générée par Sclust n'est pas la même.

Nous pouvons aussi obtenir les résultats suivants qui mettent en évidence deux et trois classes dans les données bien que les indices ne soient pas très prononcés.

SCLUST	M1	M3	M4
6	1.22695	0.02504	0.71321
5	1.43277	0.25379	0.55251
4	1.77120	0.18457	0.52646
3	2.78372	**0.16625**	**0.60160**
2	**2.97962**	0.39780	0.49897
1	-	-	-

TAB. 12.5 – Valeurs des indices de Milligan et Cooper pour Sclust

◇ Distance L_2

Les résultats des méthodes de Milligan et Cooper appliquées aux partitions générées par les quatre méthodes de classification hiérarchiques sont donnés dans le tableau 12.6.

Les valeurs des indices obtenus pour la méthode du saut minimum indiquent la présence de deux ou trois classes et sont quant à elles présentées dans le tableau 12.7.

	M1	M2	M3	M4	M5
saut minimum	3	3	2	2	3
saut maximum	4	3	2	2	3
Centroïde	4	3	2	2	3
Ward	4	3	2	2	3

TAB. 12.6 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques

saut minimum	M1	M2	M3	M4	M5
12	13.66161	2.74619	0.00000	1.00000	0.00000
11	13.92453	2.74619	0.00000	1.00000	0.00000
10	10.92181	1.86632	0.00493	0.96759	1.39264
9	12.31303	2.74619	0.00333	0.97988	0.00000
8	14.03303	2.74619	0.00201	0.98929	0.00000
7	13.73140	2.17208	0.00403	0.95882	1.49983
6	6.74503	3.12887	0.05898	0.66558	4.16515
5	7.39445	0.71548	0.04560	0.78571	0.66439
4	9.29076	0.32172	0.01674	0.94209	0.43369
3	**14.06922**	2.74619	0.00860	0.98370	0.00000
2	9.37432	**3.17743**	**0.00000**	**1.00000**	**3.43713**
1	-	2.89458	-	-	2.86634

TAB. 12.7 – Valeurs des indices de Milligan et Cooper pour la méthode du saut minimum

La partition en deux classes est toujours celle qui place le magasin de Londres dans un groupe et tous les autres magasins dans l'autre. La partition en trois classes est la suivante :
- e-Fashion Londres
- e-Fashion Rome et Madrid
- tous les autres magasins.

Pour les méthodes du saut maximum, du centroïde et de Ward, nous trouvons la présence de deux, trois ou quatre classes dans les données.

Notons aussi que les quatre méthodes hiérarchiques retrouvent les mêmes partitions en deux et trois classes mais pas la même partition en quatre classes.

Appliquons maintenant les différentes méthodes de Milligan et Cooper aux partitions générées par Sclust.

Les méthodes du C-index (M3) et Gamma (M4) retrouvent deux classes dans les don-

	M1	M3	M4
SCLUST	4	2	2

TAB. 12.8 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à Sclust

nées alors que l'indice de Calinski et Harabasz (M1) met en évidence quatre classes.

Les indices des méthodes de Milligan et Cooper prennent les valeurs suivantes :

SCLUST	M1	M3	M4
6	12.82068	0.01958	0.92339
5	15.55857	0.00878	0.92339
4	**16.42842**	0.01586	0.86851
3	14.06922	0.00860	0.98370
2	9.37432	**0.00000**	**1.00000**
1	-	-	-

TAB. 12.9 – Valeurs des indices de Milligan et Cooper pour Sclust

Les deux classes générées par Sclust sont les mêmes que celles trouvées par les méthodes hiérarchiques.

Regardons maintenant quelle est la partition en quatre classes.

```
Classe :   1  Cardinal :      4
==============================
"e-Fashion Toulouse"     "e-Fashion Berlin"
"e-Fashion Bruxelles"    "e-Fashion Paris 8ème"

Classe :   2  Cardinal :      1
==============================
"e-Fashion London"

Classe :   3  Cardinal :      6
==============================
"e-Fashion Paris 6ème"   "e-Fashion Lyon"

"e-Fashion Barcelona"    "e-Fashion Aix-Marseille"

"e-Fashion Milano"       "e-Fashion Paris 15ème"
```

```
Classe :   4  Cardinal :      2
================================
"e-Fashion Roma"          "e-Fashion Madrid"
```

Nous verrons plus tard comment justifier ces classifications.

◇ Distance L_1

Regardons les résultats obtenus en appliquant les méthodes de détermination du nombre de classes de Milligan et Cooper aux hiérarchies de partitions générées par les quatre méthodes de classification classiques. Les résultats sont les suivants :

	M1	M2	M3	M4	M5
saut minimum	2	2	2	2	2
saut maximum	2	2	2	2	2
Centroïde	2	2	2	2	2
Ward	2	2	2	2	2

TAB. 12.10 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques

Toutes les méthodes retrouvent deux classes dans les données et c'est à nouveau le magasin e-Fashion de Londres qui forme une classe à lui seul.

Voyons maintenant plus particulièrement les valeurs obtenues pour la méthode du centroïde.

Centroïde	M1	M2	M3	M4	M5
12	2.25465	2.74619	0.00000	1.00000	0.00000
11	2.34730	2.74619	0.00000	1.00000	0.00000
10	2.47370	2.74619	0.00000	1.00000	0.00000
9	2.63188	2.74619	0.00000	1.00000	0.00000
8	2.83279	2.74619	0.00031	0.98356	0.00000
7	3.07308	2.74619	0.00282	0.95370	0.00000
6	3.27955	1.34993	0.01409	0.87941	0.84880
5	3.60827	0.86832	0.02052	0.86111	0.67643
4	3.93820	0.82997	0.02839	0.85278	0.72239
3	4.47216	0.78588	0.02744	0.90761	0.72785
2	**4.98088**	1.12212	**-0.00000**	**1.00000**	0.98654
1	-	**1.76553**	-	-	**1.52298**

TAB. 12.11 – Valeurs des indices de Milligan et Cooper pour la méthode du centroïde

Appliquons maintenant les méthodes de Milligan et Cooper aux partitions générées par Sclust. Les résultats sont donnés dans le tableau 12.12.

	M1	M3	M4
SCLUST	2	2	2

TAB. 12.12 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à Sclust

Les trois méthodes retrouvent ici deux classes dans les données. Leurs indices prennent les valeurs reprises dans le tableau 12.13.

SCLUST	M1	M3	M4
6	2.73823	0.04145	0.68673
5	3.36401	0.03659	0.78747
4	3.92126	0.04104	0.85366
3	4.66519	0.04016	0.97147
2	**4.98088**	**0.00000**	**1.00000**
1	-	-	-

TAB. 12.13 – Valeurs des indices de Milligan et Cooper pour Sclust

La classification en deux classes est toujours celle isolant le magasin de Londres dans une classe.

Résultats obtenus avec le module DISS

◇ La distance PU-1

Les méthodes de détermination du nombre de classes de Milligan et Cooper pour les quatre méthodes de classification donnent les résultats suivants :

	M1	M2	M3	M4	M5
saut minimum	2	6	2	2	2
saut maximum	2	8	2	2	2
Centroïde	2	7	2	2	2
Ward	2	8	2	2	2

TAB. 12.14 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques

Les valeurs des indices des méthodes de Milligan et Cooper pour la méthode du saut minimum sont repris dans le tableau 12.15.

saut minimum	M1	M2	M3	M4	M5
k=8	2.78587	1.01419	0.00836	0.96071	0.60021
k=7	2.95054	0.94666	0.01009	0.94444	0.66489
k=6	2.61324	**1.28345**	0.05453	0.72727	1.04965
k=5	3.08118	2.74619	0.04987	0.75336	0.00000
k=4	3.63436	0.47552	0.04370	0.79444	0.48898
k=3	4.28833	0.63321	0.00998	0.96467	0.62562
k=2	**4.37665**	1.22730	**0.00000**	**1.00000**	**1.06747**
k=1	-	1.55978	-	-	1.33823

TAB. 12.15 – Valeurs des indices de Milligan et Cooper pour la méthode du saut minimum

Les méthodes de Calinski et Harabasz (M1), du C-index (M3), Gamma (M4) et de Beale (M5) indiquent clairement la présence de deux classes dans les données. Par contre, celles de Duda et Hart (M2) détecte une structure en six classes.

Le nombre de classes qui revient le plus souvent est deux. Toutes les méthodes hiérarchiques retrouvent la même partition en deux classes. Cette partition forme une classe composée du magasin e-fashion de Londres et une autre composée du reste des magasins.

Voici les résultats obtenus par les méthodes de détermination du nombre de classes de Milligan et Cooper appliquées aux partitions générées par Sclust.

	M1	M3	M4
SCLUST	2	2	2

TAB. 12.16 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à Sclust

Les indices des méthodes de Milligan et Cooper prennent les valeurs suivantes :

SCLUST	M1	M3	M4
k=8	2.66503	0.01939	0.92130
k=7	3.01870	0.00862	0.95294
k=6	3.31406	0.00858	0.95929
k=5	3.49940	0.02292	0.87692
k=4	3.97491	0.03325	0.82955
k=3	4.28833	0.00998	0.96467
k=2	**4.37665**	**0.00000**	**1.00000**
k=1	-	-	-

TAB. 12.17 – Valeurs des indices de Milligan et Cooper pour Sclust

Tous les indices sont unanimes quant à la présence de deux classes dans les données. La partition en deux classes obtenue par la méthode SCLUST est la même partition que celle obtenue pour les méthodes hiérarchiques.

◇ La distance PU-2

Les méthodes de détermination du nombre de classes de Milligan et Cooper pour les quatre méthodes de classification donnent les résultats suivants :

	M1	M2	M3	M4	M5
saut minimum	3	5	2	2	×
saut maximum	3	3	2	2	×
Centroïde	2	3	3	3	3
Ward	2	3	3	3	3

TAB. 12.18 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques

On constate que les méthodes retrouvent souvent une structure en trois classes et en deux classes. Les partitions en deux classes retrouvées par les différentes méthodes ne sont pas identiques. Les partitions en deux classes sont les suivantes :

- Pour les méthodes du saut minimum et de Ward :
  1. le magasin de Londres et
  2. les autres magasins.
- Pour les méthodes du saut maximum et du centroïde :
  1. les magasins de Londres, de Rome et de Madrid et
  2. les autres magasins.

Les partitions en trois classes sont quant à elles toutes identiques. La partition en trois classes est la suivante :

1. le magasin de Londres,
2. les magasins de Rome et de Madrid
3. les autres magasins.

Les valeurs des indices des méthodes de Milligan et Cooper pour la méthode du saut maximum sont reprises dans le tableau 12.19.

Les résultats obtenus par les méthodes de détermination du nombre de classes de Milligan et Cooper appliquées aux partitions générées par Sclust sont présentées dans le tableau 12.20.

Saut maximum	M1	M2	M3	M4	M5
k=8	1.54244	2.74619	0.00000	1.00000	0.00000
k=7	1.60650	2.74619	0.01039	0.96759	0.00000
k=6	1.63055	1.28703	0.06188	0.85588	0.79733
k=5	1.70384	0.75539	0.05702	0.85370	0.59932
k=4	1.72471	0.65084	0.06657	0.84861	0.60016
k=3	1.77092	**0.49762**	**0.03389**	**0.93207**	**0.53947**
k=2	**1.87640**	2.02243	0.07372	0.84722	1.25148
k=1	-	0.50316	-	-	0.57374

TAB. 12.19 – Valeurs des indices de Milligan et Cooper pour la méthode du saut minimum

	M1	M3	M4
SCLUST	3	2	2

TAB. 12.20 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à Sclust

Les indices des méthodes de Milligan et Cooper prennent les valeurs reprises dans le tableau 12.21.

SCLUST	M1	M3	M4
k=8	1.43585	0.07784	0.83796
k=7	1.54383	0.04480	0.87059
k=6	1.62855	0.04717	0.88602
k=5	1.64922	0.09364	0.78935
k=4	1.74540	0.10667	0.74513
k=3	**1.77092**	0.03389	0.93207
k=2	1.55217	**0.00000**	**1.00000**
k=1	-	-	-

TAB. 12.21 – Valeurs des indices de Milligan et Cooper pour Sclust

Les méthodes du C-index (M3) et Gamma (M4) détectent la présence de deux classes dans les données. La méthode de Calinski et Harabaz détecte quant à elles deux classes.

La partition en deux classes obtenue par la méthode SCLUST isole une nouvelle fois le magasin de Londres dans une classe.

La partition en trois classes est la suivante :

1. le magasin de Londres,

2. les magasins de Rome et de Madrid
3. les autres magasins.

Analyse

Pour la partition générée par Sclust, les distances L_1 et PU-1 donnent une structure en deux classes dans les données. La distance L_2 et celle de De Carvalho hésitent entre deux et quatre classes. La distance PU-2 hésite entre deux et trois classes.

Pour les hiérarchies de partitions générées par les quatre méthodes de classification classiques, les distances de De Carvalho, L_1 et PU-1 donnent deux classes. Notons néanmoins que la méthode de Duda et Hart repère de plus grands nombres de classes pour la distance PU-1. Les résultats fournis par la distance L_2 sont compris entre deux et quatre classes. Tandis que pour la distance PU-2, les résultats varient entre deux et trois classes.

Remarquons que pour la classification en deux classes, une classe est composée du magasin de Londres et l'autre des 12 individus restants. Pour mieux comprendre pourquoi, regardons les graphiques fournis par le logiciel Sodas (voir les figures 12.1 et 12.2). Remarquons que nous n'avons pas représenté ici toutes les variables sélectionnées. Les variables "étiquette article" et "étiquette couleur" ne sont pas représentables graphiquement puisqu'elles prennent beaucoup trop de modalités différentes et que le graphique devient alors illisible.

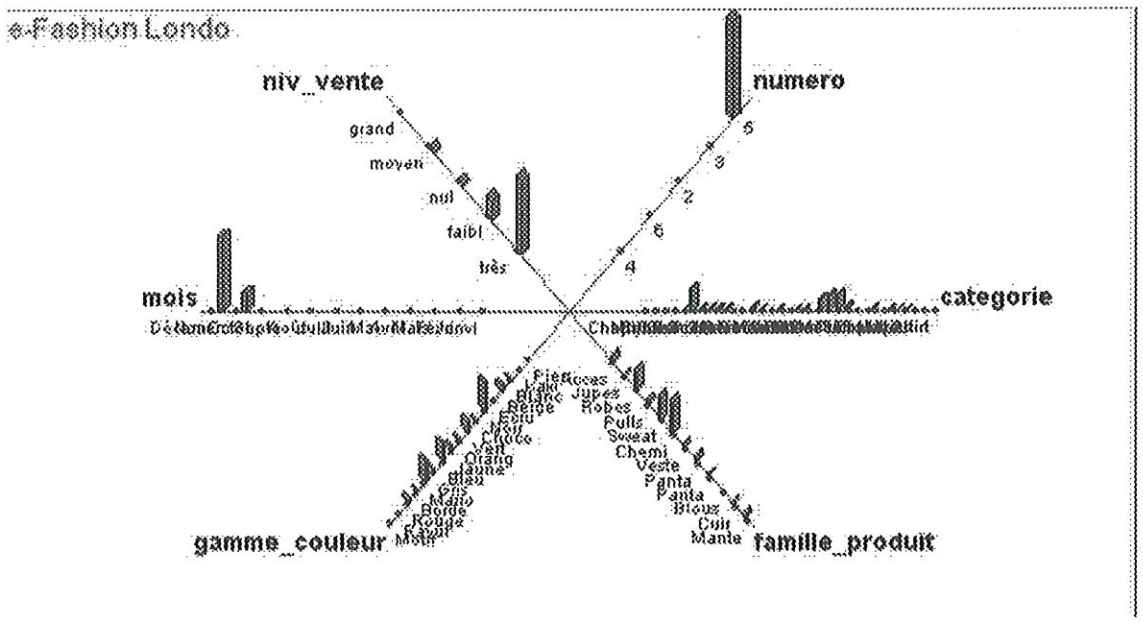


FIG. 12.1 – e-Fashion Londres

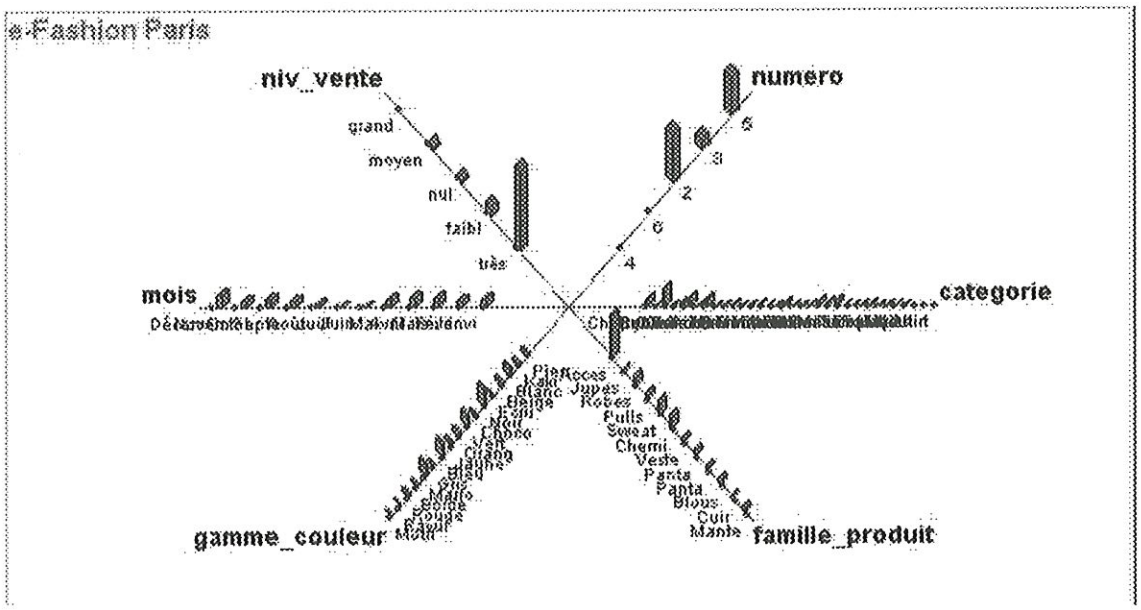


FIG. 12.2 – e-Fashion Paris 6ème

Par rapport au magasin e-Fashion Paris 6ème, nous observons que le magasin de Londres n'a utilisé qu'un type de promotion (variable "numéro"), n'a effectué des ventes qu'aux mois de novembre et décembre et n'a presque pas vendu d'accessoires mais plutôt des chemisiers, des sweats et des robes.

Lorsque nous regardons les partitions en trois classes, on constate que deux autres magasins se distinguent des autres. Ce sont les magasins de Rome et Madrid. Nous pouvons voir sur les graphiques (voir figures 12.3 et 12.4) que ces deux magasins ne vendent pour ainsi dire que des accessoires, qu'ils ont fait principalement une promotion de type 4, que leurs ventes sont réparties de la même manière tout au long de l'année et qu'ils vendent surtout des articles de couleur bleue et noire.

12.2.2 Conclusion

D'après les résultats obtenus, nous pouvons conclure soit à la présence de deux classes dans les données, la première étant composée uniquement du magasin e-Fashion de Londres et la seconde comprenant tous les magasins restants ; soit en la présence de trois classes. La troisisème classe est alors formée des magasins de Rome et Paris.

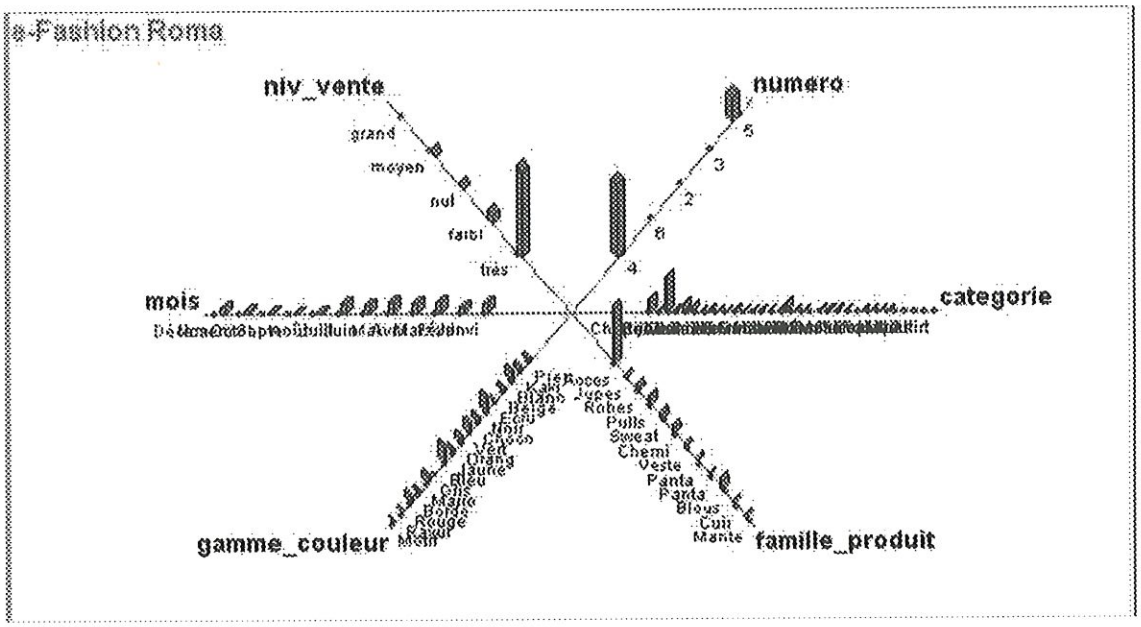


FIG. 12.3 – e-Fashion Rome

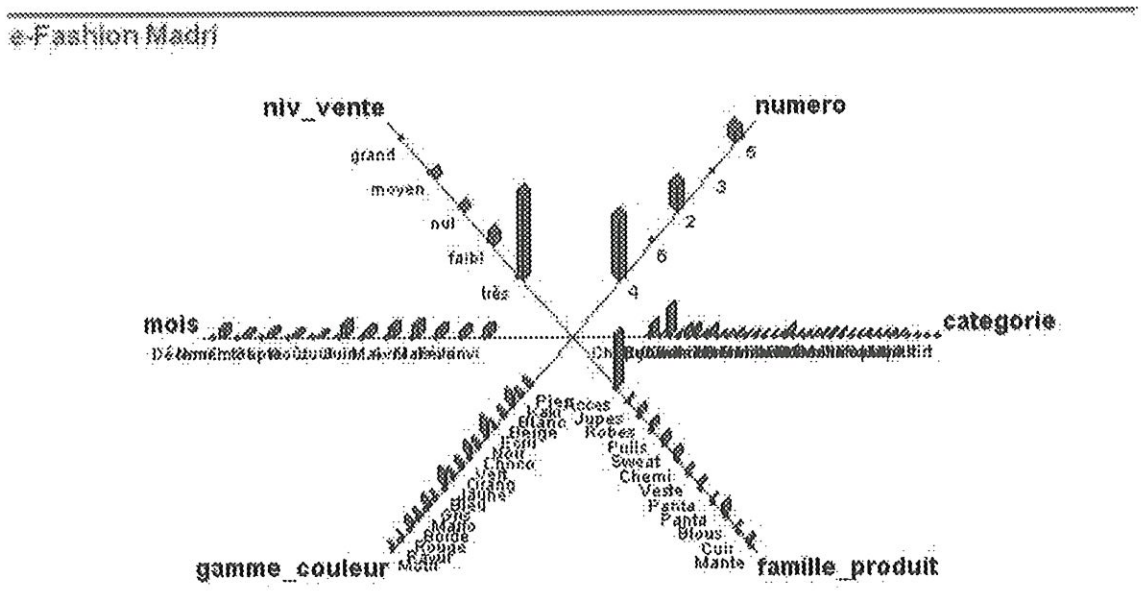


FIG. 12.4 – e-Fashion Madrid

12.2.3 Consommation [2]

Informations sur le jeu de données [2]

Ce jeu de données se trouve sur le cd de l'école Sodas de Porto. Il décrit les habitudes de consommation des ménages au Royaume-Uni et contient 25 objets symboliques décrits par 14 variables modales. Les individus représentent 25 régions du Royaume-Uni :

- "Northern metropolitan"
- "North non-metropolitan"
- "Yorks and Humberside metropolitan"
- "Yorks and Humberside non-metropolitan"
- "East Midlands non-metropolitan"
- "North West metropolitan"
- "North West non-metropolitan"
- "South East other"
- "West Midlands metropolitan"
- "West Midlands non-metropolitan"
- "East Anglia"
- "Greater London North East"
- "Greater London North West"
- "Greater London South East"
- "Greater London South West"
- "South East metropolitan"
- "South West"
- "Wales i (Gwent, 3 Glamorgans)"
- "Wales ii (Clwyd, Gwynedd, Powys, Dyfed)"
- "Scotland i (Gram, High, Tay)"
- "Scotland ii (Loth, Fife, Cen)"
- "Scotland iii metropolitan (Strathclyde)"
- "Scotland iii non-metropolitan (Strath)"
- "Scotland iv (Dum/Gall, Bord)"
- "Northern Ireland"

La première variable indique si les habitants se chauffent grâce à un chauffage central. La seconde indique le type de chauffage central qu'ils utilisent (gaz de ville, combustible solide, électricité, mazout, gaz en bouteille, combustible solide et mazout, autre). La variable suivante dit s'ils possèdent une installation au chauffage central. La quatrième variable demande si le chauffage central a été réparé durant les 12 derniers mois. Les variables suivantes concernent

- les meubles (0, de 1 à 5, de 6 à 20, plus de 20)
- le mazout (0, de 1 à 10, plus de 10)
- les tapis (0, de 1 à 5, de 6 à 20, plus de 20)
- le téléphone (0, de 1 à 5, de 6 à 10, plus de 10)

- le gaz (0, de 1 à 5, de 6 à 10, de 11 à 20, plus de 20)
- l'électricité (0, de 1 à 5, de 6 à 10, de 11 à 20, plus de 20)
- les systèmes d'égouts (0, 1, 2, 3, plus de 3)
- la licence tv (0 ou 1, 2 ou 3)
- l'eau (0, 1 ou 2, 3, 4, 5, plus de 5).

La dernière variable concerne les appareils ménagers que possèdent les foyers (réfrigérateur séparé, machine à laver, lecteur de cd, micro-onde, congélateur séparé, sèche-linge, magnétoscope, lave-vaisselle, réfrigérateur-congélateur, aucun de ceux-ci).

Notons que nous ne connaissons rien sur la classification naturelle des données.

Résultats obtenus par Séverine Collès

◇ **Distance de De Carvalho**

Les méthodes de détermination du nombre de classes de Milligan et Cooper donnent les résultats suivants pour les quatre méthodes de classification hiérarchiques :

	M1	M2	M3	M4	M5
saut minimum	3	3,6	3	3	3,6
saut maximum	4	3,6	3	3	3,6
Centroïde	3	3	3	3	3,6
Ward	×	2	3	3	3

TAB. 12.22 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques

Voici les valeurs des indices des méthodes de Milligan et Cooper pour la méthode du saut minimum :

Saut minimum	M1	M2	M3	M4	M5
k=8	14.54658	1.35041	0.05139	0.78007	1.11417
k=7	14.34058	1.48217	0.03567	0.87395	1.21390
k=6	17.92700	3.67962	0.03371	0.88414	0.00000
k=5	21.61340	**4.32994**	0.02770	0.91052	**7.98431**
k=4	26.72658	0.67673	0.00779	0.98402	0.67511
k=3	**36.11059**	3.54244	**0.00037**	**0.99944**	3.55533
k=2	10.58124	**8.19410**	0.00478	0.99124	**12.53151**
k=1	-	3.67471	-	-	3.02481

TAB. 12.23 – Valeurs des indices de Milligan et Cooper pour la méthode du saut minimum

Les méthodes de Calinski et Harabasz (M1), du C-index (M3) et Gamma (M4) indiquent clairement la présence de trois classes dans les données. Par contre, celles de Duda

et Hart (M2) et de Beale (M5) détectent une structure à deux niveaux différents de la hiérarchie. Bien qu'elles mettent en évidence une partition en six classes en atteignant leur valeur de rejet en $k = 5$, elles atteignent des valeurs fort importantes en $k = 2$ ce qui s'aligne sur les résultats obtenus par les autres méthodes.

Le nombre de classes qui revient le plus souvent est trois. Toutes les méthodes hiérarchiques retrouvent la même partition en trois classes.
La partition en trois classes est la suivante :

- Northern Ireland
- les cinq régions de Scotland
- toutes les autres régions.

La partition en six classes est donnée par :

- Greater London South West
- Northern Ireland
- Scotland iv (dum/gall, bord)
- Scotland i et iii non-metropolitan
- Scotland ii et iii metropolitan
- toutes les autres régions.

Nous regarderons plus tard si cette partition est logique ou non.

Voici les résultats obtenus par les méthodes de détermination du nombre de classes de Milligan et Cooper appliquées aux partitions générées par Sclust.

	M1	M3	M4
SCLUST	3	5	4,5

TAB. 12.24 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à Sclust

Les indices des méthodes de Milligan et Cooper prennent les valeurs reprises dans le tableau 12.25.

L'indice de Calinski et Harabasz (M1) atteint un maximum en trois classes. La méthode du C-index atteint un minimum en $k = 5$ et la méthode Gamma prend deux valeurs maximales en $k = 4$ et $k = 5$.

SCLUST	M1	M3	M4
k=8	24.85787	0.03725	0.67755
k=7	13.01632	0.08131	0.57066
k=6	9.58830	0.11372	0.56170
k=5	8.89829	**0.08384**	**0.74487**
k=4	14.31491	0.10600	**0.74487**
k=3	**14.96634**	0.09364	0.70197
k=2	13.52756	0.17179	0.67913
k=1	-	-	-

TAB. 12.25 – Valeurs des indices de Milligan et Cooper pour Sclust

Regardons plus en détail les partitions en trois, quatre et cinq classes. La partition en trois classes est la suivante :

```
Classe :   1  Cardinal :      1
================================
( 24) "Northern ireland"                    [0.0]

Classe :   2  Cardinal :     17
================================
(  0) "Northern metropolitan"               [1.0]
(  1) "North non-metropolitan"              [0.7]
(  2) "Yorks and humberside metropoli"      [0.9]
(  3) "Yorks and humberside non-metro"      [0.1]
(  4) "East midlands non-metropolitan"      [0.0]
(  5) "North west metropolitan"             [1.0]
(  6) "North west non-metropolitan"         [0.5]
(  8) "West midlands metropolitan"          [2.7]
(  9) "West midlands non-metropolitan"      [0.1]
( 10) "East anglia"                         [0.0]
( 11) "Greater london north east"           [1.7]
( 12) "Greater london north west"           [2.2]
( 13) "Greater london south east"           [2.1]
( 14) "Greater london south west"           [2.1]
( 15) "South east metropolitan"             [0.0]
( 17) "Wales i (gwent, 3 glamorgans)"       [1.2]
( 18) "Wales ii (clw, gwy, pow, dyf)"       [0.5]
```

```
Classe :   3  Cardinal :      7
================================
(  7) "South east other"                  [0.0]
( 16) "South west"                        [0.0]
( 19) "Scotland i (gram, high, tay)"      [1.2]
( 20) "Scotland ii (loth, fife, cen)"     [1.5]
( 21) "Scotland iii met (strathclyde)"    [1.5]
( 22) "Scotland iii non-met (strath)"     [1.2]
( 23) "Scotland iv (dum/gall, bord)"      [1.6]
```

La partition en quatre classes est la suivante :

```
Classe :   1  Cardinal :     13
================================
(  0) "Northern metropolitan"             [0.7]
(  1) "North non-metropolitan"            [0.4]
(  2) "Yorks and humberside metropoli"    [0.5]
(  4) "East midlands non-metropolitan"    [0.0]
(  5) "North west metropolitan"           [0.6]
(  6) "North west non-metropolitan"       [0.3]
(  8) "West midlands metropolitan"        [2.4]
( 11) "Greater london north east"         [1.2]
( 12) "Greater london north west"         [2.0]
( 13) "Greater london south east"         [1.8]
( 14) "Greater london south west"         [2.1]
( 17) "Wales i (gwent, 3 glamorgans)"     [0.8]
( 18) "Wales ii (clw, gwy, pow, dyf)"     [0.3]

Classe :   2  Cardinal :      6
================================
( 19) "Scotland i (gram, high, tay)"      [0.1]
( 20) "Scotland ii (loth, fife, cen)"     [0.8]
( 21) "Scotland iii met (strathclyde)"    [0.8]
( 22) "Scotland iii non-met (strath)"     [0.0]
( 23) "Scotland iv (dum/gall, bord)"      [0.7]
( 24) "Northern ireland"                  [3.6]

Classe :   3  Cardinal :      3
================================
(  3) "Yorks and humberside non-metro"     [3.0]
(  7) "South east other"                  [0.0]
( 16) "South west"                        [0.0]
```

```
Classe :   4  Cardinal :      3
================================
(  9) "West midlands non-metropolitan"     [3.0]
( 10) "East anglia"                        [0.0]
( 15) "South east metropolitan"            [0.0]
```

Par contre, la partition en cinq classes fournie par Sclust n'est en fait qu'une partition en trois classes puisque deux des classes ne contiennent aucun élément. Cette partition correspond à la partition en quatre classes où les classes 1 et 4 ont été regroupées.

◇ **Distance L_2**

Regardons les résultats obtenus par les méthodes de Milligan et Cooper sur les hiérarchies de partitions générées par les quatre méthodes de classification classiques. Les résultats sont les suivants :

	M1	M2	M3	M4	M5
saut minimum	3	3	5	5	3
saut maximum	4	2,4	5	5	2,4
Centroïde	3	3	5	5	3
Ward	2	2	4	4	2

TAB. 12.26 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques

Notons tout d'abord que les quatre méthodes de classification classiques donnent toutes des hiérarchies de partitions différentes. Nous n'allons détailler ici que les résultats obtenus avec la méthode du saut maximum.

Saut maximum	M1	M2	M3	M4	M5
k=8	14.27375	2.18835	0.01778	0.82715	1.74080
k=7	15.39926	3.67962	0.01890	0.81946	0.00000
k=6	16.62893	1.55936	0.02602	0.77966	1.22903
k=5	16.07007	2.34098	**0.00554**	**0.98156**	1.90418
k=4	**17.53602**	3.23728	0.00818	0.97868	2.25095
k=3	11.29891	**5.50747**	0.04169	0.90502	**5.83815**
k=2	16.90273	2.79801	0.04690	0.89275	1.92962
k=1	-	**5.15354**	-	-	4.83191

TAB. 12.27 – Valeurs des indices de Milligan et Cooper pour la méthode du saut maximum

La partition en cinq classes retrouvée par les méthodes du C-index (M3) et Gamma (M4) pour la méthode du saut maximum est la suivante :
- Northern Ireland
- Scotland iv (dum/gall, bord)
- Wales ii (clw, gwy, pow, dyf) et Scotland i
- Les trois autres régions de Scotland
- Toutes les autres régions.

La partition en deux classes retrouvée par la méthode de Beale est la suivante :
- Wales ii, Scotland i, Scotland iv et Northern Ireland
- Toutes les autres régions.

Par rapport aux résultats obtenus avec la distance de De Carvalho, la partition en trois classes revient moins souvent. Les différentes méthodes de détermination du nombre de classes hésitent entre deux, trois, quatre et cinq classes. Remarquons aussi qu'elles n'indiquent pas de partition en trois classes pour les méthodes du saut maximum et de Ward.

Notons de plus que, la plupart du temps, les méthodes hiérarchiques ne forment pas les mêmes hiérarchies de partitions avec la distance L_2 qu'avec la distance de De Carvalho.

Regardons maintenant les résultats obtenus par les méthodes de détermination du nombre de classes de Milligan et Cooper appliquées aux partitions générées par Sclust.

	M1	M3	M4
SCLUST	2	3,5	3

TAB. 12.28 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à Sclust

En effet, les indices des méthodes de Milligan et Cooper prennent les valeurs reprises dans le tableau 12.29.

SCLUST	M1	M3	M4
8	13.89174	0.02812	0.73194
7	14.43556	0.02183	0.87461
6	11.21341	0.04871	0.84937
5	16.27142	**0.01075**	0.95751
4	15.04607	0.02088	0.94977
3	20.07273	**0.01142**	**0.97684**
2	**23.42443**	0.04876	0.92494
1	-	-	-

TAB. 12.29 – Valeurs des indices de Milligan et Cooper pour Sclust

Regardons la partition en deux classes.

```
Classe :   1  Cardinal :      7
================================
( 18) "Wales ii (clw, gwy, pow & dyf)"    [1.4]
( 19) "Scotland i (gram, high & tay)"     [0.1]
( 20) "Scotland ii (loth, fife & cen)"    [0.5]
( 21) "Scotland iii met (strathclyde)"    [0.9]
( 22) "Scotland iii non-met (strath)"     [0.7]
( 23) "Scotland iv (dum/gall & bord)"     [0.7]
( 24) "Northern ireland"                  [2.7]

Classe :   2  Cardinal :     18
================================
(  0) "Northern metropolitan"             [1.7]
(  1) "North non-metropolitan"            [0.8]
(  2) "Yorks and humberside metropoli"    [0.5]
(  3) "Yorks and humberside non-metro"    [0.8]
(  4) "East midlands non-metropolitan"    [0.2]
(  5) "North west metropolitan"           [0.5]
(  6) "North west non-metropolitan"       [0.7]
(  7) "South east other"                  [1.1]
(  8) "West midlands metropolitan"        [0.8]
(  9) "West midlands non-metropolitan"    [0.5]
( 10) "East anglia"                       [2.3]
( 11) "Greater london north east"         [1.1]
( 12) "Greater london north west"         [1.0]
( 13) "Greater london south east"         [1.5]
( 14) "Greater london south west"         [1.1]
( 15) "South east metropolitan"           [0.6]
```

```
( 16) "South west"                          [1.5]
( 17) "Wales i (gwent & 3 glamorgans)"      [1.4]
```

Regardons à présent la partition en trois classes. Cette partition n'est pas la même que celle obtenue pour la distance de De Carvalho. Ici, Northern Ireland ne forme plus une classe à lui seul et les différentes régions de Scotland ne sont plus toutes dans la même classe.

```
Classe :   1  Cardinal :      3
================================
( 18) "Wales ii (clw, gwy, pow & dyf)"      [0.8]
( 23) "Scotland iv (dum/gall & bord)"       [0.9]
( 24) "Northern ireland"                    [1.3]

Classe :   2  Cardinal :     18
================================
(  0) "Northern metropolitan"               [1.7]
(  1) "North non-metropolitan"              [0.8]
(  2) "Yorks and humberside metropoli"      [0.5]
(  3) "Yorks and humberside non-metro"      [0.8]
(  4) "East midlands non-metropolitan"      [0.2]
(  5) "North west metropolitan"             [0.5]
(  6) "North west non-metropolitan"         [0.7]
(  7) "South east other"                    [1.1]
(  8) "West midlands metropolitan"          [0.8]
(  9) "West midlands non-metropolitan"      [0.5]
( 10) "East anglia"                         [2.3]
( 11) "Greater london north east"           [1.1]
( 12) "Greater london north west"           [1.0]
( 13) "Greater london south east"           [1.5]
( 14) "Greater london south west"           [1.1]
( 15) "South east metropolitan"             [0.6]
( 16) "South west"                          [1.5]
( 17) "Wales i (gwent & 3 glamorgans)"      [1.4]

Classe :   3  Cardinal :      4
================================
( 19) "Scotland i (gram, high & tay)"       [1.7]
( 20) "Scotland ii (loth, fife & cen)"      [0.3]
( 21) "Scotland iii met (strathclyde)"      [1.4]
( 22) "Scotland iii non-met (strath)"       [0.6]
```

Et pour terminer, regardons la partition en cinq classes.

```
Classe :   1  Cardinal :     17
================================
(  0) "Northern metropolitan"              [1.9]
(  1) "North non-metropolitan"             [0.9]
(  2) "Yorks and humberside metropoli"     [0.5]
(  3) "Yorks and humberside non-metro"     [0.9]
(  4) "East midlands non-metropolitan"     [0.3]
(  5) "North west metropolitan"            [0.5]
(  6) "North west non-metropolitan"        [0.7]
(  7) "South east other"                   [1.3]
(  8) "West midlands metropolitan"         [0.8]
(  9) "West midlands non-metropolitan"     [0.6]
( 11) "Greater london north east"          [1.1]
( 12) "Greater london north west"          [1.1]
( 13) "Greater london south east"          [1.6]
( 14) "Greater london south west"          [1.1]
( 15) "South east metropolitan"            [0.7]
( 16) "South west"                         [1.7]
( 17) "Wales i (gwent & 3 glamorgans)"     [1.5]

Classe :   2  Cardinal :      2
================================
( 10) "East anglia"                        [1.0]
( 18) "Wales ii (clw, gwy, pow & dyf)"     [1.0]

Classe :   3  Cardinal :      3
================================
( 20) "Scotland ii (loth, fife & cen)"     [0.6]
( 21) "Scotland iii met (strathclyde)"     [1.7]
( 22) "Scotland iii non-met (strath)"      [0.8]

Classe :   4  Cardinal :      1
================================
( 24) "Northern ireland"                   [0.0]

Classe :   5  Cardinal :      2
================================
( 19) "Scotland i (gram, high & tay)"      [1.0]
( 23) "Scotland iv (dum/gall & bord)"      [1.0]
```

◇ **Distance L_1**

Appliquons les méthodes de détermination du nombre de classes de Milligan et Cooper aux hiérarchies de partitions générées par les quatre méthodes de classification classiques. Voici les résultats obtenus :

	M1	M2	M3	M4	M5
saut minimum	5	5	5	2	5
saut maximum	2	2,4	4	4	2
Centroïde	4	5	5	5	4
Ward	2	2,5	3	3	2

TAB. 12.30 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques

Regardons les résultats obtenus pour la méthode de Ward qui sont les suivants :

Ward	M1	M2	M3	M4	M5
k=8	3.73933	3.67962	0.05288	0.82469	0.00000
k=7	3.89299	1.72860	0.05563	0.80977	1.22778
k=6	4.16950	1.35534	0.05376	0.82272	1.02274
k=5	4.61519	1.04175	0.06500	0.79068	0.84437
k=4	4.75804	**2.86479**	0.08520	0.75442	1.55247
k=3	5.34879	1.29667	**0.03057**	**0.93773**	1.08079
k=2	**7.32395**	1.98761	0.06497	0.87639	1.45554
k=1	-	**2.67201**	-	-	**2.09366**

TAB. 12.31 – Valeurs des indices de Milligan et Cooper pour la méthode de Ward

La méthode de Calinski et Harabasz (M1) retrouve deux classes dans les données comme l'indice de Beale (M5). La méthode de Duda et Hart (M2) hésite quant à elle entre deux et cinq classes et les méthodes du C-index (M3) et Gamma (M4) indiquent trois classes.

La partition en cinq classes est la suivante :

1. Northern Ireland
2. Wales ii et Scotland iv
3. les quatre autres régions de Scotland
4. North non-metropolitan, Yorks and Humberside non-metropolitan, East Midlands non-metropolitan, South east other, West Midlands non-metropolitan, East anglia, South West

5. les autres régions.

Nous obtenons la partition en trois classes en regroupant les classes 1 et 2 et les classes 4 et 5.

Le premier groupe de la partition en deux classes reprend les classes 1, 2, 4 et 5 de la partition en cinq classes et le deuxième groupe correspond à la classe 3.

Alors qu'en utilisant la distance de De Carvalho, la plupart des méthodes de détermination du nombre de classes indiquaient trois classes, avec la distance L_1, seules les méthodes du C-index (M3) et Gamma (M4) appliquées à la hiérarchie de partitions de Ward mettent en évidence trois classes.

Les méthodes de détermination du nombre de classes appliquées aux partitions générées par Sclust indiquent deux, trois ou six classes présentes dans les données comme le montre le tableau suivant :

	M1	M3	M4
SCLUST	2	3	6

TAB. 12.32 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à Sclust

En effet, les indices des méthodes de Milligan et Cooper prennent les valeurs suivantes :

SCLUST	M1	M3	M4
k=8	3.55806	0.05989	0.80307
k=7	3.55247	0.04319	0.93022
k=6	4.13831	0.05443	**0.95813**
k=5	4.04670	0.09781	0.93258
k=4	4.83627	0.08079	0.93634
k=3	5.26865	**0.03248**	0.90988
k=2	**7.32395**	0.06497	0.87922
k=1	-	-	-

TAB. 12.33 – Valeurs des indices de Milligan et Cooper pour Sclust

La partition en deux classes retrouvée par Sclust est la même que celle obtenue pour la distance L_2.

La partition en trois classes est la suivante.

```
Classe :   1  Cardinal :     18
==============================
(  0) "Northern metropolitan"              [1.4]
(  1) "North non-metropolitan"             [0.9]
(  2) "Yorks and humberside metropoli"     [0.7]
(  3) "Yorks and humberside non-metro"     [1.0]
(  4) "East midlands non-metropolitan"     [0.5]
(  5) "North west metropolitan"            [0.8]
(  6) "North west non-metropolitan"        [0.9]
(  7) "South east other"                   [0.9]
(  8) "West midlands metropolitan"         [0.9]
(  9) "West midlands non-metropolitan"     [0.7]
( 10) "East anglia"                        [1.4]
( 11) "Greater london north east"          [1.2]
( 12) "Greater london north west"          [1.1]
( 13) "Greater london south east"          [1.2]
( 14) "Greater london south west"          [1.2]
( 15) "South east metropolitan"            [0.8]
( 16) "South west"                         [1.1]
( 17) "Wales i (gwent & 3 glamorgans)"     [1.2]

Classe :   2  Cardinal :      1
==============================
( 24) "Northern ireland"                   [0.0]

Classe :   3  Cardinal :      6
==============================
( 18) "Wales ii (clw, gwy, pow & dyf)"     [1.4]
( 19) "Scotland i (gram, high & tay)"      [0.6]
( 20) "Scotland ii (loth, fife & cen)"     [0.8]
( 21) "Scotland iii met (strathclyde)"     [1.0]
( 22) "Scotland iii non-met (strath)"      [0.8]
( 23) "Scotland iv (dum/gall & bord)"      [1.3]
```

Comme pour la partition en trois classes pour la distance de De Carvalho, Northern Ireland forme une classe à lui seul et les cinq régions de Scotland sont regroupées dans la même classe mais au lieu d'être associées avec South East other et South West, elles sont associées à Wales ii.

Et pour terminer, regardons la partition en six classes.

```
Classe :   1  Cardinal :      4
================================
( 19) "Scotland i (gram, high & tay)"     [1.2]
( 20) "Scotland ii (loth, fife & cen)"    [0.6]
( 21) "Scotland iii met (strathclyde)"    [1.3]
( 22) "Scotland iii non-met (strath)"     [0.9]
}
Classe :   2  Cardinal :      7
================================
(  3) "Yorks and humberside non-metro"     [0.9]
(  4) "East midlands non-metropolitan"    [0.8]
(  7) "South east other"                  [0.9]
(  9) "West midlands non-metropolitan"    [0.9]
( 14) "Greater london south west"         [1.4]
( 15) "South east metropolitan"           [1.0]
( 16) "South west"                        [1.0]

Classe :   3  Cardinal :      1
================================
( 10) "East anglia"                       [0.0]

Classe :   4  Cardinal :      1
================================
( 24) "Northern ireland"                  [0.0]

Classe :   5  Cardinal :      2
================================
( 18) "Wales ii (clw, gwy, pow & dyf)"    [1.0]
( 23) "Scotland iv (dum/gall & bord)"     [1.0]

Classe :   6  Cardinal :     10
================================
(  0) "Northern metropolitan"             [1.3]
(  1) "North non-metropolitan"            [1.0]
(  2) "Yorks and humberside metropoli"    [0.9]
(  5) "North west metropolitan"           [0.6]
(  6) "North west non-metropolitan"       [0.9]
(  8) "West midlands metropolitan"        [0.8]
( 11) "Greater london north east"         [1.0]
( 12) "Greater london north west"         [1.1]
```

```
( 13) "Greater london south east"          [1.1]
( 17) "Wales i (gwent & 3 glamorgans)"     [1.3]
```

Résultats obtenus avec le module DISS

◇ Distance PU-1

Appliquons les méthodes de détermination du nombre de classes de Milligan et Cooper aux hiérarchies de partitions générées par les quatre méthodes de classification classiques. Voici les résultats obtenus :

	M1	M2	M3	M4	M5
saut minimum	3	3	4	4	3
saut maximum	2	2	3	3	2
Centroïde	3	3,5	4	4	3
Ward	2	2,6	3	3	2

TAB. 12.34 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques

En ce qui concerne les méthodes hiérarchiques, on constate que les méthodes de Milligan et Cooper hésitent entre deux, trois et quatre classes.

Regardons les résultats obtenus pour la méthode du centroïde qui sont les suivants :

Centroïde	M1	M2	M3	M4	M5
k=8	4.30563	3.17870	0.02582	0.90081	2.11513
k=7	4.61830	0.70632	0.02770	0.89560	0.66087
k=6	4.64940	1.11140	0.01759	0.95659	0.94948
k=5	5.25173	**0.50646**	0.00402	0.99340	0.56626
k=4	6.01532	2.44259	**0.00230**	**0.99704**	1.70668
k=3	**6.91501**	**1.26385**	0.02445	0.94150	**1.05910**
k=2	3.48317	3.30547	0.02356	0.97494	2.68632
k=1	-	1.17280	-	-	0.99572

TAB. 12.35 – Valeurs des indices de Milligan et Cooper pour la méthode du centroïde

Les méthode de Calinski et Harabasz (M1) et de Beale (M5) retrouvent trois classes dans les données. La méthode de Duda et Hart (M2) hésite quant à elle entre trois et cinq classes et les méthodes du C-index (M3) et Gamma (M4) indiquent quatre classes.

La partition en trois classes est la suivante :

1. Northern Ireland

2. les cinq régions de Scotland
3. les autres régions.

Nous obtenons la partition en quatre classes en formant une classe contenant la région wales ii.

Tout comme pour la distance de De Carvalho, la plupart des méthodes de détermination du nombre de classes indiquent trois classes, avec la distance PU-1. Néanmoins des structures en deux et quatre classes sont aussi retrouvées.

Notons que les méthodes de Ward et du saut maximum retrouvent toujours les mêmes partitions. Il en est de même pour les méthodes du saut minimum et du centroïde.

Les méthodes de détermination du nombre de classes appliquées aux partitions générées par Sclust indiquent deux ou trois classes dans les données comme le montre le tableau suivant :

	M1	M3	M4
SCLUST	2	3	3

TAB. 12.36 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à Sclust

En effet, les indices des méthodes de Milligan et Cooper prennent les valeurs suivantes :

SCLUST	M1	M3	M4
k=7	4.31592	0.08271	0.69429
k=6	4.78366	0.05967	0.82044
k=5	5.55492	0.06085	0.83661
k=4	6.11589	0.07810	0.86082
k=3	7.23475	**0.01348**	**0.98109**
k=2	**9.95896**	0.03823	0.92759
k=1	-	-	-

TAB. 12.37 – Valeurs des indices de Milligan et Cooper pour Sclust

La partition en deux classes et celle en trois classes retrouvées par Sclust sont les mêmes que celles obtenue pour la distance L_2.

◇ Distance PU-2

Appliquons les méthodes de détermination du nombre de classes de Milligan et Cooper aux hiérarchies de partitions générées par les quatre méthodes de classification classiques. Voici les résultats obtenus :

	M1	M2	M3	M4	M5
saut minimum	3	3,5	4	4	3,5
saut maximum	2	6	2	2	2,4
Centroïde	3	3,5	4	4	3,5
Ward	2	6	2	2	2,4

TAB. 12.38 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques

On constate une nouvelle fois que les méthodes de Milligan et Cooper hésitent entre deux, trois et quatre classes.

Regardons les résultats obtenus pour la méthode du saut maximum qui sont les suivants :

Saut maximum	M1	M2	M3	M4	M5
k=8	2.02653	1.11179	0.12704	0.80102	0.84655
k=7	2.08716	3.67962	0.13531	0.78688	0.00000
k=6	2.19094	**1.67991**	0.11251	0.81892	1.11184
k=5	2.28491	2.32813	0.13432	0.76947	0.96613
k=4	2.49312	0.78216	0.16575	0.69471	**0.68621**
k=3	2.67369	1.55659	0.22291	0.57407	1.08876
k=2	**3.51104**	0.66876	**0.02774**	**0.92219**	**0.66554**
k=1	-	1.18525	-	-	1.00369

TAB. 12.39 – Valeurs des indices de Milligan et Cooper pour la méthode du saut maximum

Les méthodes de Calinski et Harabasz (M1), du C-index (M3) et Gamma (M4) retrouvent deux classes dans les données. Les méthodes de Duda et Hart (M2) et de Beale (M5) hésitent entre deux et quatre classes.

La partition en deux classes est la suivante :

1. Northern Ireland, les cinq régions de Scotland et Wales ii,
2. les autres régions.

La partition en quatre classes est celle-ci :

1. Northern Ireland, Scotland iv et Wales ii,

2. les autres Scotland,
3. North non-metropolitan, Yorks and Humberside non-metropolitan, East midlands non-metropolitan, South east other, West midlands non-metropolitan, South west,
4. les autres régions.

Les méthodes de détermination du nombre de classes appliquées aux partitions générées par Sclust indiquent deux ou trois classes dans les données comme le montre le tableau suivant :

	M1	M3	M4
SCLUST	2	3	3

TAB. 12.40 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées à Sclust

En effet, les indices des méthodes de Milligan et Cooper prennent les valeurs suivantes :

SCLUST	M1	M3	M4
k=7	2.04583	0.16998	0.71492
k=6	2.19699	0.12129	0.82693
k=5	2.30891	0.10374	0.85523
k=4	2.63277	0.12155	0.87525
k=3	2.74185	**0.01169**	**0.97542**
k=2	**3.51104**	0.02774	0.92411
k=1	-	-	-

TAB. 12.41 – Valeurs des indices de Milligan et Cooper pour Sclust

La partition en deux classes et celle en trois classes retrouvées par Sclust sont les mêmes que celles obtenues pour la distance PU-1.

Analyse

Les méthodes de détermination du nombre de classes indiquent la présence de une à six classes dans les données. Alors que pour la distance de De Carvalho, il semble y avoir trois classes (ou peut-être six), pour les distances L_2, L_1, PU-1 et PU-2, cela varie de deux à six.

Notons que les hiérarchies de partitions sont aussi fort changeantes selon la distance et les méthodes hiérarchiques utilisées. Cependant, en général, nous remarquons que Northern Ireland se distingue des autres régions ainsi que Scotland iv et Wales ii.

Regardons donc les Zoom Star en trois dimensions associés à ces individus.

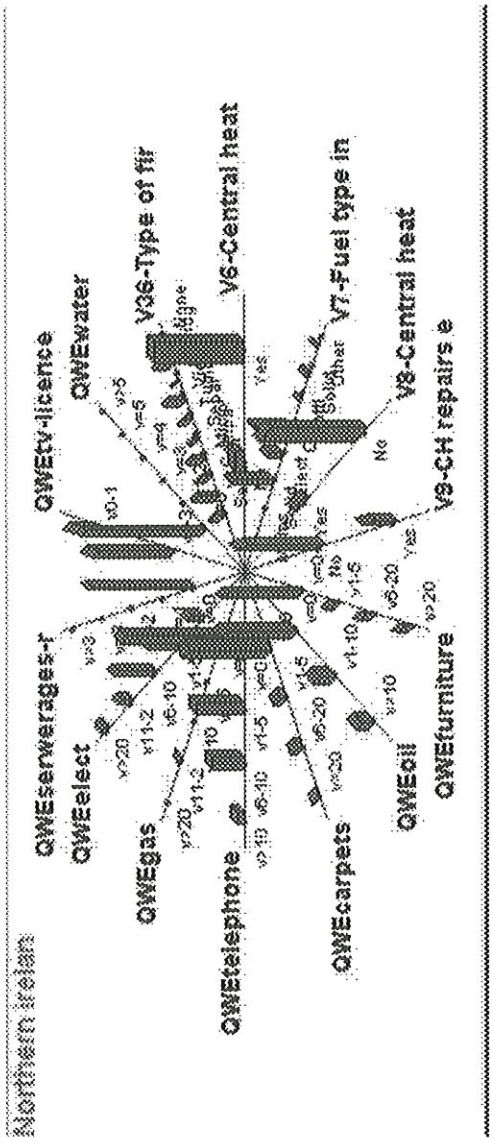


FIG. 12.5 – Northern Ireland

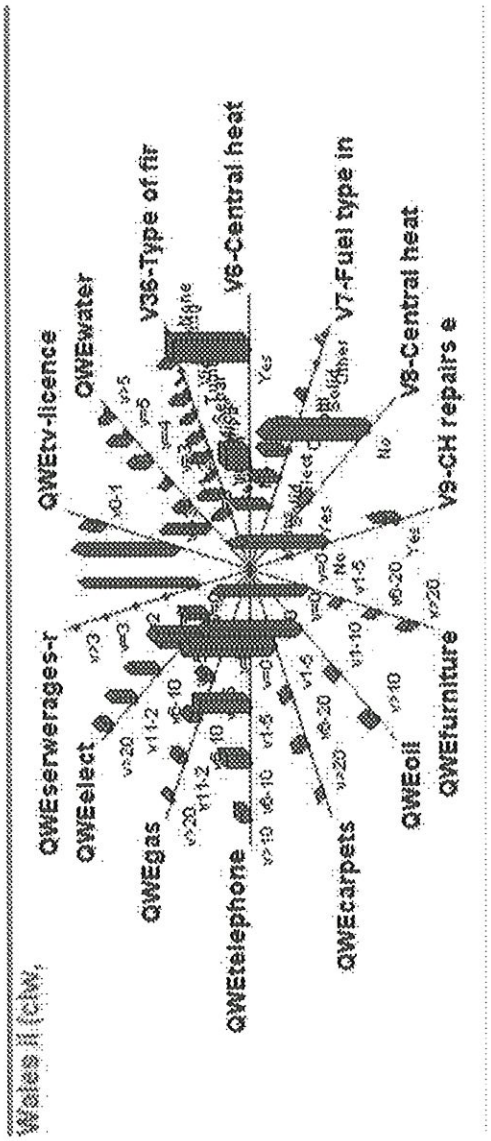


FIG. 12.6 – Wales ii

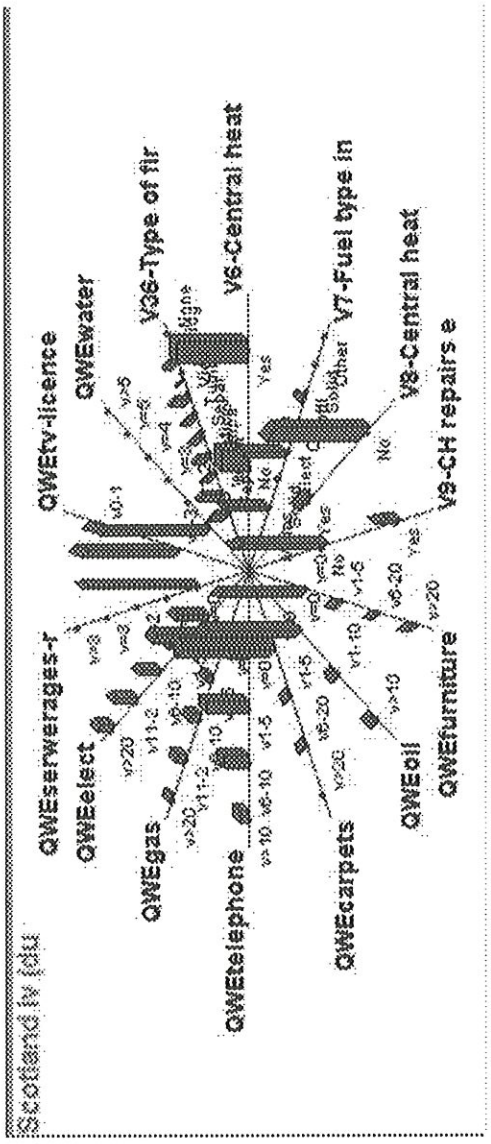


FIG. 12.7 – Scotland iv

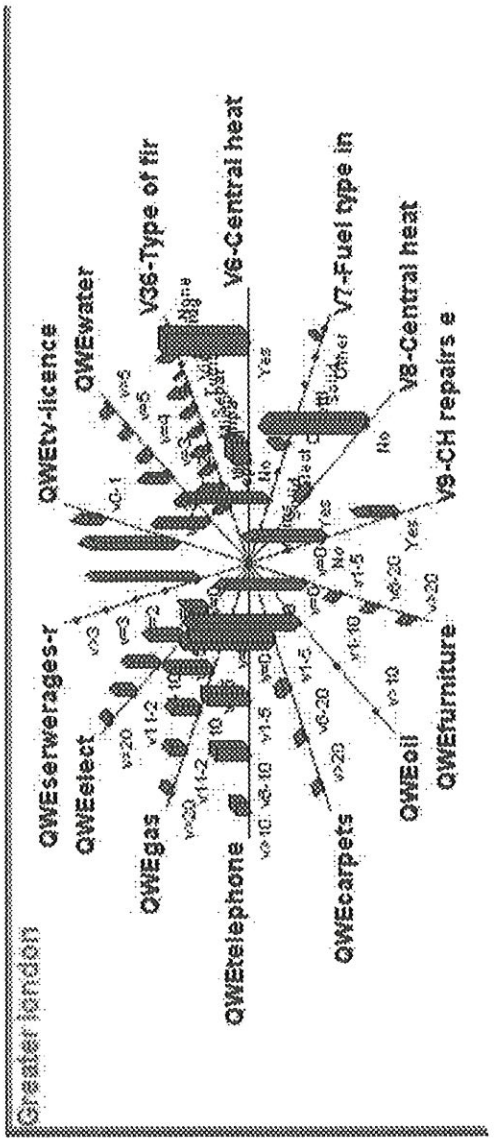


FIG. 12.8 – Greater London South East

Chapitre 13

Combinaisons de variables

13.1 Introduction

Ce chapitre est le plus important du mémoire car il est la concrétisation de ce à quoi nous voulions arriver. Le jeu de données que nous étudions est décrit pour deux types de variables. Nous montrons qu'il nous est possible d'obtenir une détermination du nombre de classes pour les différentes méthodes de classification.

Le jeu de données est décrit par deux types de variable, des variables de type intervalle et multivaluées. Dans les chapitres traitant ces types de variable, nous avons remarqué qu'une seule des distances donnait des résultats concluants. Nous étudierons donc la détermination du nombre de classe pour cette distance qui est le distance U-2.

13.2 Voitures

13.2.1 Informations sur le jeu de données

Ce jeu de données se trouve dans le logiciel SODAS. Il décrit les caractéristiques de 32 voitures. Les individus sont les suivants :

- Alfa 145,
- Alfa 156,
- Alfa 166,
- Aston Martin,
- Audi A3,
- Audi A6,
- Audi A8,
- Bmw serie 3,
- Bmw serie 5,
- Bmw serie 7,

- Ferrari,
- Punto,
- Fiesta,
- Focus,
- Honda NSK,
- Lamborghini,
- Lancia Y,
- Lancia K,
- Maserati GT,
- Mercedes SL,
- Mercedes Classe C,
- Mercedes Classe E,
- Mercedes Classe S,
- Nissan Micra,
- Corsa,
- Vectra,
- Porsche,
- Twingo,
- Rover 25,
- Rover 75,
- Skoda Fabia,
- Skoda Octavia,
- Passat.

Les onze variables sont :
- huit variables de type intervalle :
  1. la pression,
  2. la cylindrée,
  3. la vitesse maximale,
  4. l'accélération,
  5. passo (en italien, nous n'avons pas trouver la définition de ce mot),
  6. la longueur,
  7. la largeur et
  8. la hauteur.
- trois variables multivaluées :
  9. l'alimentation qui possède deux modalités : l'essence et le diesel.
  10. la traction qui possède trois modalités : avant, arrière et intégrale.
  11. la catégorie : utilitaire, berline, "grande voiture sportive" et sportive.

Notons que nous ne connaissons rien sur la classification naturelle des données.

13.2.2 Résultats obtenus avec le module DISS

Pour rappel, nous avons utilisé uniquement la distance U-2.

Voici les résultats obtenus pour la détermination du nombre de classes appliquées aux partitions générées par les quatre méthodes hiérarchiques :

	M1	M2	M3	M4	M5
saut minimum	5,3	5	5	5	5
saut maximum	2	4,2	3	3	2
Centroïde	4	3	?,3	?,3	3
Ward	2	4,2	3	3	2

TAB. 13.1 – Résultats obtenus par les méthodes de Milligan et Cooper appliquées aux hiérarchies de partitions générées par quatre méthodes de classification classiques

On constate une nouvelle fois que les méthodes de Milligan et Cooper hésitent pour déterminer la structure présente dans ce jeu de données.

Les cinq méthodes de Milligan et Cooper associée à la partition du saut minimum détectent cinq classes dans les données.

La partition en cinq classes est la suivante :

1. Alfa 145, Alfa 156, Alfa 166, Audi A3, Audi A6, Audi A8, Bmw serie 3, Bmw serie 5, Punto, Fiesta, Focus, Lancia Y, Lancia K, Maserati GT,Mercedes Classe C, Nissan Micra, Corsa, Vectra,Twingo, Rover 25, Rover 75, Skoda Fabia, Skoda Octavia et Passat ;
2. Aston Martin, Ferrari et Porsche ;
3. Bmw serie 7, Honda NSK et Mercedes SL ;
4. Lamborghini ;
5. Mercedes Classe E et Mercedes Classe S.

Pour la partition générée pour le saut maximum et la méthode de Ward, les méthodes de Milligan et Cooper préconise deux ou trois classes dans les données.

Les partitions en deux classes sont équivalentes et sont :

1. Aston Martin, Audi A8, Bmw serie 7, Ferrari, Honda NSK, Lamborghini, Maserati GT, Mercedes SL, Mercedes Classe E, Mercedes Classe S et Porsche ;
2. les autres voitures.

```
(  0) Alfa  145            [1.0]   ( 11) Punto              [0.6]
( 12) Fiesta               [1.0]   ( 13) Focus              [1.3]
( 16) Lancia Y             [0.9]   ( 23) Nissan Micra       [1.2]
( 24) Corsa                [1.0]   ( 27) Twingo             [1.3]
( 28) Rover 25             [0.6]   ( 30) Skoda Fabia        [0.6]
( 31) Skoda Octavia        [1.5]

Classe :   2  Cardinal :      7
================================

(  3) Aston Martin         [1.1]   ( 10) Ferrari            [0.7]
( 14) Honda  NSK           [0.8]   ( 15) Lamborghini        [2.3]
( 18) Maserati GT          [0.5]   ( 19) Mercedes SL        [0.8]
( 26) Porsche              [0.7]

Classe :   3  Cardinal :      5
================================

(  6) Audi A8              [1.3]   (  8) Bmw serie 5        [0.9]
(  9) Bmw serie 7          [0.7]   ( 21) Mercedes Classe E  [1.0]
( 22) Mercedes Classe S    [1.1]

Classe :   4  Cardinal :     10
================================

(  1) Alfa  156            [0.5]   (  2) Alfa  166          [0.9]
(  4) Audi A3              [1.1]   (  5) Audi A6            [1.2]
(  7) Bmw serie 3          [0.8]   ( 17) Lancia K           [0.9]
( 20) Mercedes Classe C    [1.1]   ( 25) Vectra             [1.4]
( 29) Rover 75             [0.7]   ( 32) Passat             [1.3]
```

Voyons finalement la partition en six classes est celle-ci :

```
Classe :   1  Cardinal :      5
================================

(  0) Alfa  145            [1.3]   ( 13) Focus              [0.9]
( 25) Vectra               [0.3]   ( 31) Skoda Octavia      [0.8]
( 32) Passat               [1.7]

Classe :   2  Cardinal :      8
```

Quant aux partitions en trois classes, elles voient une séparation des voitures sportives. Les classes se distinguent surtout par le biais de la variable pression.

La partition en quatre classes forme une classe contenant uniquement la Lamborghini. Cette voiture se distingue des autres au niveau de la variable pression. Cette variable de type intervalle est dans ce cas de petite longueur et possède des valeurs extrêmes relativement élevées par rapport aux autres individus.

La partition en cinq classes préconisée par la méthode du saut minimum nous semble cependant incorrecte à première vue. Alors que quatre des classes sont formées avec des voitures sportives, elle met la Maserati GT dans la classe contenat des berlines et des utilitaires.

Pour la méthode non-hiérarchique SCLUST, les méthodes de Milligan et Cooper hésitent entre quatre et six classes.

La partition en quatre classes générée par SCLUST différencie les classes grâce à la variable de type intervalle cylindrée et à la variable modale catégorie.

La diversité des partitions obtenues mettent en évidence des déterminations du nombre de classes assez différentes. Toutes les partitions obtenues semblent néanmoins assez correctes et aucune ne nous semble meilleure qu'une autre.

Bibliographie

[1] S. DELOGNE, Méthodes de détermination du nombre de classes pour des données symboliques de type intervalles. FUNDP,2002.

[2] S. COLLES, Méthodes de détermination du nombre de classes pour des données symboliques multivaluées et modales. FUNDP, 2003.

[3] H.-H. BOCK et E. DIDAY, Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data. Springer-Verlag,2000.

[4] G.W. MILLIGAN et M.C. COOPER, An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50(2),159-179, Juin 1985.

[5] B. EVERITT, Cluster Analysis, Arnold, London, 1993.

[6] J.L. CHANDON et S. PINSON, Analyse typologique : théorie et applications, Masson, Paris, 1981.

[7] G. CELEUX, E. DIDAY, G. GOVAERT, Y. LECHEVALLIER, et H. RALANBONDRAINY, Classification automatique des données, DUNOD-Informatique, Bordas, 1989.

[8] T. CALINSKI et J. HARABASZ, A dendrite method for cluster analysis, Communications in Statistics, 3, 1-27, 1974

[9] R.O. DUDA et P.E. HART, Pattern classification and scene analysis, Wiley-Interscience, New-York, 1973.

[10] L.J. HUBERT et J.R. LEVIN, A general statistical framework for assessing categorical clustering in free recall, Psychological Bulletin, 83(6), 1072-1080, 1976.

[11] F.B. BAKER et L.J. HUBERT, Measuring the power of hierarchical cluster analysis, Journal of the American Statistical Association, 70, 31-38, 1975.

[12] E.M.L. BEALE, Cluster Analysis, London, Scientific Control Systems, 1969.

[13] L. BREIMAN, J.H. FRIEDMAN, R.A. OLSHEN et C.J. STONE, Classification and Regression Trees, Belmont Editions, 1984.

[14] ASSO, Analysis System of Symbolic Official Data, Diss-Vdiss,http ://www.info.fundp.ac.be/asso/assoprivate/Dec2003/Tutorial/DISS_VDISS_tutorial.pdf, 2003

[15] ASSO, Analysis System of Symbolic Official Data, DSTAT Help Guide, http ://www.info.fundp.ac.be/asso/assoprivate/Dec2003/HelpGuide/TM_1_2_helpguide.pdf, 2003

[16] M. TOUATI et E. DIDAY , Logiciel SODAS,http ://www.ceremade.dauphine.fr/ touati/sodas-pagegarde.htm, Lise-Ceremade

[17] ASSO, Téléchargement logiciel SODAS, http ://www-rocq.inria.fr/sodas/WP1/asso-parser/