MASTER EN SCIENCES MATHÉMATIQUES

Etude comparative de quatre méthodes de classification symboliques et applications

Kindermans, Sigried

Award date:
2005

FUNDP
Faculté des Sciences
Département de Mathématique

Rempart de la Vierge, 8
B-5000 Namur Belgique

Etude comparative de quatre méthodes de classification symboliques et applications

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

Sigried KINDERMANS

Promoteur : Prof. André HARDY

Année Académique 2004-2005

Mes plus sincères remerciements s'adressent en premier lieu à André Hardy, promoteur de ce mémoire, pour son aide, sa disponibilité et sa gentillesse.

Je remercie également le Docteur Jamart qui m'a permis de travailler dans un domaine qui m'intéresse particulièrement : la médecine. Ses conseils m'ont fortement aidé.

Je remercie aussi Anne Lemaitre qui m'a aidée durant mon séjour Erasmus à Marseille.

À Stéphane qui m'a soutenu et encouragé tout au long de mes études.

Enfin, je n'oublie pas ma famille et mon entourage.

Merci à vous tous !

Résumé

La classification automatique occupe une place importante en analyse des données. Le problème consiste en la décomposition d'un ensemble d'individus, décrits par des variables, en un certain nombre de classes homogènes. Dans ce mémoire, nous nous intéressons aux individus caractérisés par des variables symboliques. Quatre méthodes de classification symboliques sont présentées. Ces méthodes sont ensuite testées sur des ensembles de données générés aléatoirement. Enfin, deux applications sont exposées. La première traite d'un problème rencontré en logopédie. La seconde concerne les pays membres de l'Union Européenne.

Abstract

Clustering (or unsupervised classification) is an important part in data analysis. The purpose is to split up a set of individuals -described by variables- in a number of homogeneous clusters. In this paper we only consider individuals that are characterized by symbolic variables. Four symbolic clustering methods are presented. These methods are tested afterwards on some artificial data. Two applications are eventually exposed. The first is dealing with a medical problem. The second one treats on the European Union.

Table des matières

Introduction générale 1

1 Données symboliques 2
- 1.1 Présentation . . . 3
- 1.2 Types de variables symboliques . . . 3
  - 1.2.1 Variables intervalles . . . 3
  - 1.2.2 Variables multivaluées . . . 4
  - 1.2.3 Variables modales . . . 5
- 1.3 Tableau de données symboliques . . . 6
- 1.4 Dissimilarités entre objets symboliques . . . 7
  - 1.4.1 Variables intervalles . . . 7
  - 1.4.2 Variables multivaluées . . . 8
  - 1.4.3 Variables modales . . . 10
- 1.5 Des données classiques aux données symboliques . . . 11

2 Classification automatique 15
- 2.1 Formulation mathématique du problème . . . 16
- 2.2 Structures classificatoires . . . 17
  - 2.2.1 Les partitions . . . 17
  - 2.2.2 Les hiérarchies . . . 18
- 2.3 Méthodes de classification . . . 19
  - 2.3.1 Les méthodes hiérarchiques . . . 19
  - 2.3.2 Les méthodes de partitionnement . . . 20
- 2.4 Méthode des nuées dynamiques . . . 21
  - 2.4.1 La notion d'inertie . . . 21
  - 2.4.2 Théorème de Huygens . . . 21
  - 2.4.3 Inerties associées à une partition . . . 22
  - 2.4.4 La méthode des nuées dynamiques . . . 23

3 Méthodes de classification symboliques **27**
3.1 La méthode SCLUST 28
3.1.1 Présentation 28
3.1.2 Principe général 28
3.1.3 Algorithme 31
3.2 La méthode DIV 32
3.2.1 Présentation 32
3.2.2 Extension du critère de la variance intra-classe 32
3.2.3 Bipartition d'une classe 33
3.2.4 Choix de la classe à diviser 36
3.2.5 Algorithme 36
3.2.6 Résultats de l'algorithme 37
3.3 La méthode SCLASS 37
3.3.1 Le processus de Poisson non-homogène 37
3.3.2 Hypothèse générale 38
3.3.3 La méthode des noyaux 38
3.3.4 Recherche de "bosses" et test de la multi-modalité 39
3.3.5 La règle de coupure 39
3.3.6 L'élagage 40
3.3.7 Application aux données de type intervalle 41
3.3.8 Algorithme 42
3.3.9 Résultats de l'algorithme 42
3.4 La méthode HIPYR 43
3.4.1 Présentation 43
3.4.2 Principe général 43
3.4.3 Algorithme 45

4 Comparaison des méthodes sur des données artificielles **46**
4.1 Données avec deux classes hypersphériques 47
4.1.1 Le jeu de données 47
4.1.2 La méthode SCLUST 47
4.1.3 La méthode DIV 49
4.1.4 La méthode SCLASS 50
4.1.5 La méthode HIPYR 52
4.1.6 Conclusion 53
4.2 Données avec trois classes hypersphériques 54
4.2.1 Le jeu de données 54
4.2.2 La méthode SCLUST 55
4.2.3 La méthode DIV 56

4.2.4 La méthode SCLASS 58
4.2.5 La méthode HIPYR 62
4.2.6 Conclusion 63
4.3 Données avec deux classes allongées 64
4.3.1 Le jeu de données 64
4.3.2 La méthode SCLUST 65
4.3.3 La méthode DIV 67
4.3.4 La méthode SCLASS 69
4.3.5 La méthode HIPYR 71
4.3.6 Conclusion 72
4.4 Données avec deux classes emboîtées 73
4.4.1 Le jeu de données 73
4.4.2 La méthode SCLUST 73
4.4.3 La méthode DIV 75
4.4.4 La méthode SCLASS 76
4.4.5 La méthode HIPYR 77
4.4.6 Conclusion 78
4.5 Bilan 79

5 Applications **80**
5.1 Application 1 : Le *Voice Handicap Index* 81
5.1.1 *Voice Handicap Index* (VHI) 81
5.1.2 *Voice Handicap Index* adapté à la voix chantée 82
5.1.3 Contexte de l'étude 84
5.1.4 Motivation de l'étude 86
5.1.5 Données symboliques 86
5.1.6 Classification des sujets 87
5.2 Application 2 : Union Européenne 93
5.2.1 Le jeu de données 94
5.2.2 Classification des pays sur base de toutes les variables 96
5.2.3 Classification des pays sur base des indicateurs économiques 105
5.2.4 Classification des pays sur base des indicateurs démographiques . . . 110
5.2.5 Conclusion de l'étude 114

Conclusion générale **115**

Annexes **116**

Bibliographie **124**

Introduction générale

Les progrès incessants de l'informatique ont permis le recueil, la gestion et le traitement de données en quantité de plus en plus importantes dans divers domaines de l'activité humaine comme la biologie, la médecine ou encore l'économie. Ceci a privilégié le développement des données dites symboliques, souvent utilisées pour résumer de grands ensembles de données sans perte d'information.

Nous nous intéressons dans ce mémoire au problème de la classification automatique, qui occupe une place importante en analyse des données. Il consiste en la décomposition d'un ensemble d'individus, décrits par des variables, en un certain nombre de groupes homogènes. De nombreuses méthodes permettent la résolution de ce type de problème. L'objectif de ce mémoire est de comparer les méthodes de classification symboliques incluses dans le logiciel SODAS 2 développé dans le cadre du projet européen ASSO (Analysis System of Symbolic Official data).

Le premier chapitre sera consacré à la présentation des données symboliques. Nous aborderons dans un second chapitre le problème de la classification automatique. Nous y décrirons deux grandes familles de méthodes permettant la résolution de ce problème. Ensuite, nous présenterons quatre méthodes de classification symboliques incluses dans le logiciel SODAS 2. Nous testerons ces méthodes sur des ensembles de données générés de manière à mettre en évidence des structures variées. Finalement, nous exposerons deux applications. La première traite d'un problème rencontré en logopédie. La seconde concerne les pays membres de l'Union Européenne.

Chapitre 1

Données symboliques

Nous consacrons ce premier chapitre aux données symboliques. Notre objectif est de familiariser le lecteur avec le nouveau concept de données symboliques.

Nous commencerons par définir trois types de données symboliques, que nous illustrerons par un exemple. Ensuite, nous montrerons comment les données symboliques peuvent être stockées dans une matrice de manière à pouvoir être traitées. Puis, nous définirons les dissimilarités entre objets symboliques. Finalement, nous illustrerons comment transformer une matrice de données classiques en une matrice de données symboliques. Nous montrerons ce qu'une telle transformation engendre au niveau des données.

1.1 Présentation

En analyse de données symboliques, l'ensemble des objets E peut être défini de deux façons différentes :

1. Un ensemble $E = \Omega = \{x_1,x_2,\ldots,x_n\}$ d'individus appelés **objets du premier ordre** ;
2. Un ensemble $E = \{C_1,C_2,\ldots\}$ de classes $C_i \subseteq \Omega$ d'individus appelées **objets du second ordre**.

Une **variable symbolique** Y d'espace d'observation $\mathcal{Y}$ est définie de la manière suivante :

$$\begin{array}{rccl} Y: & E & \rightarrow & \mathcal{B} \quad \forall\, x_k \in E \\ & x_k & \rightsquigarrow & Y(x_k) \end{array}$$

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

Dans ce chapitre, nous définissons trois types de variables :
- les variables intervalles ;
- les variables multivaluées ;
- les variables modales.

1.2 Types de variables symboliques

1.2.1 Variables intervalles

Soit $E = \{x_1,x_2,\ldots,x_n\}$ un ensemble de n objets symboliques.

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

Dans ce cas, $\mathcal{B} = \mathcal{P}(\mathcal{Y})$ est l'ensemble des intervalles fermés bornés de $\mathbb{R}$.

Exemple 1.2.1 - Variable intervalle

Soient
- $E = \{$habitants d'une commune$\}$;
- $Y =$ le temps consacré à l'exercice d'un ou plusieurs sports (en heures/semaine) ;
- $\mathcal{B} = \{[a,b] \mid a,b \in \mathbb{R}^+, 0 \leq a \leq b < \infty\}$.

Nous pouvons avoir, par exemple, $Y(x_k) = [1,2]$ et $Y(x_l) = [2,5]$ comme valeurs de la variable Y pour les objets $x_k,x_l \in E$.

◇

1.2.2 Variables multivaluées

Soit $E = \{x_1, x_2, \ldots, x_n\}$ un ensemble de n objets symboliques.

Une variable Y est dite **multivaluée** lorsque les valeurs $Y(x_k)$ sont toutes des sous-ensembles finis du domaine $\mathcal{Y}$, c'est-à-dire $|Y(x_k)| < \infty, \forall\, x_k \in E$.

Il existe deux types de variables multivaluées :
- les variables multivaluées catégoriques ;
- les variables multivaluées quantitatives.

Une variable est dite **multivaluée catégorique** si $\mathcal{Y}$ a un nombre fini de catégories de manière à ce que $|Y(x_k)| < \infty, \forall\, x_k \in E$.

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

Exemple 1.2.2 - Variables multivaluées catégorique et quantitative

Considérons
- un ensemble E de domaines skiables des Alpes françaises
  $E = \{$Les Trois Vallées, Espace Killy, Paradiski$\}$;
- Y_1 = les stations de ski composant chacun de ces domaines
  $\mathcal{Y}_1 = \{$Les Ménuires, Val Thorens, Les Arcs, Val d'Isère, Tignes, $\ldots\}$;
- Y_2 = le nombre de pistes rouge des deux plus grosses stations de chaque domaine
  $\mathcal{Y}_2 = \mathbb{N}_0^+$.

Nous avons pour les trois domaines skiables de l'ensemble E :

Domaines skiables	Y_1	Y_2
Les Trois Vallées	{Les Ménuires, Val Thorens, Méribel, Courchevel, La Tania}	{35,29}
Espace Killy	{Val d'Isère, Tignes}	{18,17}
Paradiski	{Les Arcs, Peisey Vallandry, La Plagne}	{34,32}

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

◇

1.2.3 Variables modales

Une variable **modale** Y sur un ensemble $E = \{x_1,x_2,\ldots,x_n\}$ d'objets dont l'espace d'observation est $\mathcal{Y}$, est une fonction

$$Y(x_k) = \left(U(x_k),\pi_{x_k}\right) \quad \forall\, x_k \in E$$

où

- π_{x_k} est une mesure ou une distribution (fréquence, probabilité, poids) sur les valeurs possibles de $\mathcal{Y}$;
- $U(x_k) \subseteq \mathcal{Y}$ est le support de π_{x_k} dans le domaine $\mathcal{Y}$.

Exemple 1.2.3 - Variable modale

Soient

- $E = \{x_1,x_2,\ldots,x_{100}\}$ un ensemble de 100 travailleurs ;
- $\widetilde{Y}$ = la durée du trajet pour se rendre au travail (en minutes) ;
- C_1 = la classe des travailleurs qui arrivent au travail le matin parmi les 10 premiers.

Les valeurs prises par la variable $\widetilde{Y}$ pour la classe C_1 sont les suivantes : 40, 65, 35, 25, 45, 30, 60, 20, 25 et 40.

La variable modale qui décrit la durée du trajet dans la classe C_1 peut avoir une réalisation sous la forme d'un histogramme :

$$Y(C_1) = \left\{\left(]15,25],\frac{3}{10}\right),\left(]25,35],\frac{2}{10}\right),\left(]35,45],\frac{3}{10}\right),\left(]45,55],\frac{0}{10}\right),\left(]55,65],\frac{2}{10}\right)\right\}.$$

◇

1.3 Tableau de données symboliques

Considérons un ensemble E de n objets symboliques sur lesquels nous avons mesuré p variables symboliques $Y_1, Y_2, \ldots, Y_p$ où Y_j a pour espace d'observation $\mathcal{Y}_j$.

Les observations ainsi obtenues constituent une matrice $\underline{\mathrm{X}}$ de données symboliques ayant n lignes et p colonnes

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

où x_{ij} est la valeur prise par la variable symbolique Y_j pour l'objet x_i.

La cellule x_{ij} de la matrice de données symboliques peut contenir des ensembles, des intervalles ou encore des histogrammes.

La ligne i de la matrice correspond à la **description symbolique** de l'objet i sur les p variables.

Exemple 1.3.1 - Tableau de données symboliques

Considérons

- E = {Londres, Paris, Athènes} un ensemble de trois grandes villes européennes ;
- Y_1 = le nombre d'habitants (minimum et maximum entre les années 1999 et 2004),
  $\mathcal{B}_1 = \{[\alpha,\beta] = [\min, \max] \,|\, 0 \leq \alpha \leq \beta < \infty\}$,
  Y_1 est une variable intervalle ;
- Y_2 = les deux sites les plus visités de chaque ville,
  $\mathcal{Y}_2$ = {Tour Eiffel, Musée Grévin, Big Ben, ... },
  $\mathcal{B}_2 = \mathcal{P}(Y_2)$,
  Y_2 est une variable multivaluée catégorique ;
- Y_3 = les moyens de transport utilisés dans une ville ainsi que le pourcentage de personnes empruntant chacun de ces différents moyens de transport,
  $\mathcal{Y}_3$ = {Voiture, Taxi, Bus, Métro},
  $\mathcal{B}_3$ est donc l'ensemble des distributions de fréquences sur $\mathcal{Y}_3$,
  Y_3 est une variable modale.

Le tableau de données symboliques est le suivant:

Villes	Y_1 (10^3)	Y_2	Y_3
Londres	[2.332,2.387]	{Big Ben, Tower Bridge}	{V 0.2; T 0.3; B 0.2; M 0.3}
Paris	[2.133,2.212]	{Tour Eiffel, Arc de Triomphe}	{V 0.1; T 0.1; B 0.3; M 0.5}
Athènes	[742,791]	{Parthénon, Stade olympique}	{V 0.3; T 0.2; B 0.2; M 0.3}

La deuxième ligne de ce tableau

$$x_2' = \Big([2.133,2.212],\{\text{Tour Eiffel, Arc de Triomphe}\},\{\text{V}\,0.1;\text{T}\,0.1;\text{B}\,0.3;\text{M}\,0.5\}\Big)$$

correspond à la description symbolique de la ville de Paris.

◇

1.4 Dissimilarités entre objets symboliques

1.4.1 Variables intervalles

Considérons une matrice de données symboliques $\underline{\text{X}}$ composée de n objets symboliques sur lesquels nous avons mesuré p variables intervalles $Y_1,Y_2,\ldots,Y_p$.

Nous avons

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

où $x_{kj} = Y_j(x_k) = [\alpha_{kj},\beta_{kj}]$ est la valeur de la variable Y_j pour l'objet $x_k \in E$.

Nous définissons une mesure de dissimilarité sur l'ensemble des objets E à partir de p indices de dissimilarité sur les $\mathcal{B}_j$. Ainsi,

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

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

1. La distance de Hausdorff définie par

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

   consiste à prendre le maximum entre la valeur absolue de la différence des bornes inférieures des deux intervalles et la différence de leurs bornes supérieures ;

2. La distance $\mathcal{L}_1$ définie par

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

   est la somme des valeurs absolues des différences entre les bornes inférieures et supérieures des deux intervalles ;

3. La distance $\mathcal{L}_2$ définie par

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

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

Pour définir une mesure de dissimilarité sur l'ensemble des objets E, nous combinons les p indices de dissimilarité définis sur les $\mathcal{B}_j$. Ainsi, nous définissons

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

où δ_j est une des mesures de dissimilarité définies précédemment.

1.4.2 Variables multivaluées

Considérons un ensemble E de n objets symboliques décrits par p variables multivaluées $Y_1, Y_2, \ldots, Y_p$ dont les espaces d'observation sont respectivement $\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_p$.

Soient
- $Y_j(x_k)$ un ensemble de catégories et $\mathcal{B}_j = \mathcal{P}(\mathcal{Y}_j)$;
- m_j le nombre de catégories prises par $\mathcal{Y}_j$;
- $q_{j,x_k}(c_s)$ la fréquence associée à la catégorie c_s $(s = 1, \ldots m_j)$ de la variable Y_j pour l'objet x_k

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

La description symbolique de l'objet $x_k \in E$ est donnée par

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

La matrice de données symboliques $\underline{X}$ est ainsi transformée en une matrice de fréquences

$$\widetilde{X} = \begin{pmatrix} q_{1,x_1}(c_1) & \cdots & q_{1,x_1}(c_{m_1}) & \cdots & q_{p,x_1}(c_1) & \cdots & q_{p,x_1}(c_{m_p}) \\ q_{1,x_2}(c_1) & \cdots & q_{1,x_2}(c_{m_1}) & \cdots & q_{p,x_2}(c_1) & \cdots & q_{p,x_2}(c_{m_p}) \\ \vdots & \cdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ q_{1,x_n}(c_1) & \cdots & q_{1,x_n}(c_{m_1}) & \cdots & q_{p,x_n}(c_1) & \cdots & q_{p,x_n}(c_{m_p}) \end{pmatrix}$$

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

Tout comme pour les variables intervalles, nous définissons une mesure de dissimilarité sur l'ensemble des objets E à partir des indices de dissimilarité définis sur les ensembles $\mathcal{B}_j$.

Ainsi,

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

À partir de deux ensembles de catégories x_{kj} et x_{lj}, nous pouvons définir les trois distances suivantes :

1. La distance $\mathcal{L}_1$ définie par

$$\delta_j(x_{kj},x_{lj}) = \sum_{i=1}^{|\mathcal{Y}_j|} \left| q_{j,x_k}(c_i) - q_{j,x_l}(c_i) \right|,$$

consiste à prendre la somme sur $i = 1,\ldots,|\mathcal{Y}_j|$ des valeurs absolues des différences entre les fréquences associées aux catégories c_i de Y_j;

2. La distance $\mathcal{L}_2$ définie par

$$\delta_j(x_{kj},x_{lj}) = \sum_{i=1}^{|\mathcal{Y}_j|} \left(q_{j,x_k}(c_i) - q_{j,x_l}(c_i) \right)^2,$$

est la somme sur $i = 1,\ldots,|\mathcal{Y}_j|$ des carrés des différences entre les fréquences associées aux catégories c_i de Y_j;

3. La distance de Carvalo définie par

$$\delta_j(x_{kj},x_{lj}) = \sum_{i=1}^{|y_j|} \left(\gamma\, q_{j,x_k}(c_i) + \gamma'\, q_{j,x_l}(c_i)\right),$$

où

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

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

Pour nous ramener à une mesure dissimilarité sur l'ensemble E des objets, nous définissons

$$\begin{aligned} d \;&: E \times E \;\rightarrow\; \mathbb{R}^+ \\ (x_k,x_l) \;&\rightsquigarrow\; d(x_k,x_l) = \left(\sum_{j=1}^{p} \delta_j^2\,(x_{kj},x_{lj})\right)^{\frac{1}{2}} \end{aligned}$$

où δ_j est une des mesures de dissimilarité définies précédemment.

1.4.3 Variables modales

Le cas des variables modales est semblable à celui des variables multivaluées. Il suffit simplement de remplacer les fréquences $q_{j,x_k}(c_s)$ par les valeurs de la distribution π_{j,x_k} associées à chacune des catégories de $Y_j(x_k)$.

1.5 Des données classiques aux données symboliques

Nous présentons un exemple de transformation d'une matrice de données classiques en une matrice de données symboliques. Cet exemple nous permettra notamment d'illustrer la manière dont les données classiques peuvent être réduites et résumées.

Nous utilisons des données extraites d'un rapport publié en 2001 par la Banque Mondiale et les Nations Unies. Par souci de simplification, nous ne prendrons en compte que 20 pays parmi les 156 ayant fait l'objet de cette étude.

Les variables mesurées sur les différents pays sont les suivantes :

- $\widetilde{Y}_1 =$ `Richesse` correspond à un indice de richesse économique. Les valeurs prises par cette variable nous renseigne sur le niveau de richesse économique du pays.

  Cette variable est une variable qualitative nominale à 6 modalités :
  - TFA : Pays à niveau de richesse économique très faible,
  - FAI : Pays à niveau de richesse économique faible,
  - MFA : Pays à niveau de richesse économique moyennement faible,
  - MEL : Pays à niveau de richesse économique moyennement élevé,
  - ENM : Pays à niveau de richesse économique élevé non membres de l'Organisation du Commerce et du Développement Économique,
  - EME : Pays à niveau de richesse économique élevé membres de l'Organisation du Commerce et du Développement Économique ;

- $\widetilde{Y}_2 =$ `Développement` correspond à un indice de développement économique. Les différentes valeurs prises par cette variable nous renseignent sur le niveau de développement économique du pays.

  Cette variable est une variable qualitative nominale à 3 modalités :
  - PVD : Pays en voie de développement,
  - PEM : Pays émergent,
  - PDE : Pays développé ;

- $\widetilde{Y}_3 =$ `Continent` correspond au continent du pays. Il s'agit d'une variable qualitative nominale à 5 modalités :
  - Amérique,
  - Europe,
  - Afrique,
  - Océanie,
  - Asie ;

- $\widetilde{Y}_4$ = Produit National Brut par habitant (en USD);
- $\widetilde{Y}_5$ = Taux d'exportation (en % du PNB);
- $\widetilde{Y}_6$ = Taux d'importation (en % du PNB);
- $\widetilde{Y}_7$ = Taux d'inflation (en %).

Les variables $\widetilde{Y}_4$,$\widetilde{Y}_5$,$\widetilde{Y}_6$ et $\widetilde{Y}_7$ sont des variables quantitatives continues.

Les données sont reprises dans le tableau 1.1. Chaque ligne de ce tableau correspond à un pays. Chaque pays est décrit par les 7 facteurs économiques présentés précédemment.

Pays	Rich.	Dévelop.	Continent	PNB/hab. (USD)	Export. (% PNB)	Import. (% PNB)	Inflation
Japon	EME	PDE	Asie	35.620	9,98	8,44	-0,59
Etats-Unis	EME	PDE	Amérique	34.100	10,72	13,47	2,21
Belgique	EME	PDE	Europe	24.540	88,05	84,65	1,22
France	EME	PDE	Europe	23.810	28,66	27,25	0,92
Chypre	ENM	PDE	Europe	12.370	48,21	45,23	2,87
Israël	ENM	PDE	Asie	16.710	39,99	46,94	1,71
Albanie	FAI	PVD	Europe	1.120	18,89	40,39	-1,19
Sénégal	FAI	PVD	Afrique	490	30,54	39,56	0,73
Kenya	FAI	PVD	Afrique	350	26,49	35,63	6,82
Arabie Saoudite	MEL	PVD	Asie	7.230	49,57	25,73	16,08
Croatie	MEL	PEM	Europe	4.620	45,01	50,63	6,45
Brésil	MEL	PEM	Amérique	3.590	10,81	12,42	8,23
Argentine	MEL	PVD	Amérique	7.480	10,81	11,45	0,77
Maroc	MFA	PVD	Afrique	1.180	31,22	37,36	1,55
Mexique	MFA	PEM	Amérique	5.110	31,06	33,01	11,98
Afrique du Sud	MFA	PVD	Afrique	3.060	28,59	25,67	6,99
Chine	MFA	PEM	Asie	840	25,89	23,21	0,92
Thaïlande	MFA	PVD	Asie	2.010	66,97	58,92	1,17
Madagascar	TFA	PVD	Afrique	250	24,64	34,78	7,1
Zambie	TFA	PVD	Afrique	310	17,28	39,68	30,05

TAB. 1.1 – *Matrice de données classiques.*

Reprenons à présent les 20 pays et montrons comment nous pouvons passer de la matrice de données classiques (tableau 1.1) à une matrice de données symboliques. Pour ce faire, nous avons utilisé le module DB2SO du logiciel SODAS.

Nous allons agréger les individus en 6 classes $C_1, \ldots, C_6$ tel que
- C_1 = {Pays à niveau de richesse économique élevé membres de l'OCDE} ;
- C_2 = {Pays à niveau de richesse économique élevé non membres de l'OCDE} ;
- C_3 = {Pays à niveau de richesse économique faible} ;
- C_4 = {Pays à niveau de richesse économique moyennement élevé} ;
- C_5 = {Pays à niveau de richesse économique faible} ;
- C_6 = {Pays à niveau de richesse économique très faible}.

L'ensemble $E = \{C_1, \ldots, C_6\}$ de classes $C_i \subseteq \Omega$ d'individus appelées objets du second ordre peut être décrit par la matrice de données symboliques (tableau 1.2).

Chaque classe de ce tableau est décrite par 6 variables symboliques $Y_1, \ldots, Y_6$ de telle façon que $Y_j(C_i)$ caractérise l'ensemble $\{\widetilde{Y}_j(x_k) \mid x_k \in C_i \subseteq \mathcal{Y}_j\}$.

Les variables `Développement` et `Continent` sont des variables modales. Les autres variables sont de type intervalle. Les bornes inférieure et supérieure d'un intervalle sont respectivement les valeurs minimale et maximale prises par la variable classique correspondante.

En analysant le tableau de données symboliques 1.2, nous pouvons d'ores et déjà dégager des tendances générales en ce qui concerne l'information contenue dans les données initiales.

En effet, au vu de ce tableau :
- les pays à niveau de richesse économique élevé membres de l'OCDE sont pour 50% Européens, 25% Américains et 25% Asiatiques. Le Produit National Brut par habitant de ces pays est élevé. Il est compris entre 23.810 et 35.620 USD;
- parmi les pays à niveau de richesse économique moyennement élevé, 50% d'entre eux sont en voie de développement et 50% sont émergents;
- les pays à niveau de richesse économique très faible sont des pays africains. Le Produit National Brut par habitant varie entre 250 et 310 USD pour ces pays. Ce sont des pays en voie de développement.

Nous avons pu constater au travers de cet exemple que l'approche classique n'est pas la même que l'approche symbolique. En transformant la matrice de données classiques en une matrice de données symboliques, nous réduisons les données en agrégeant des individus en classes. Cette réduction de données fait en sorte que nous obtenons une vue d'ensemble des données de départ.

	Développement	Continent	PNB/habitant (USD)	Exportation (% PNB)	Importation (% PNB)	Inflation
EME	PDE (1,00)	Europe (0,50), Amérique (0,25), Asie (0,25)	[23.810 : 35.620]	[9,98 : 88,05]	[8,44 : 84,65]	[−0,59 : 2.21]
ENM	PDE (1,00)	Asie (0,50), Europe (0,50)	[12.370 : 16.710]	[39,99 : 48,21]	[45,23 : 46,94]	[1,71 : 2,87]
FAI	PVD (1,00)	Afrique (0,67), Europe (0,33)	[350 : 1.120]	[18,89 : 30,54]	[35,63 : 40,39]	[−1,19 : 6,82]
MEL	PEM (0,50), PVD (0,50)	Amérique (0,50), Asie (0,25), Europe (0,25)	[3.590 : 7.480]	[10,81 : 49,57]	[11,45 : 50,63]	[0,77 : 16,08]
MFA	PVD (0,60), PEM (0,40)	Afrique (0,40), Asie (0,40), Amérique (0,20)	[840 : 5.110]	[25,89 : 66,97]	[23,21 : 58,92]	[0,92 : 11,98]
TFA	PVD (1,00)	Afrique (1,00)	[250 : 310]	[17,28 : 24,64]	[34,78 : 39,68]	[7,10 : 30,05]

TAB. 1.2 – *Matrice de données symboliques.*

Chapitre 2

Classification automatique

La classification automatique occupe une place importante en analyse des données. Elle vise à trouver les structures intrinsèques des données en les organisant en groupes homogènes, appelés classes.

Le problème consiste en la décomposition d'un ensemble d'individus en un certain nombre de groupes de manière à ce que :

- les individus d'un même groupe soient les plus semblables possibles ;
- les individus de groupes différents soient les plus différents possibles.

Nous débuterons ce chapitre par la formalisation du problème de la classification automatique en termes mathématiques. Puis, nous rappellerons les différentes structures classificatoires que les méthodes de classification peuvent engendrer. Ensuite, nous nous intéresserons aux méthodes de classification. Nous décrirons les principales étapes des méthodes hiérarchiques et des méthodes de partitionnement. Nous terminerons par la description de la méthode de partitionnement des nuées dynamiques.

2.1 Formulation mathématique du problème

Soient
- $E = \{x_1, \ldots, x_n\}$ un ensemble de n individus ;
- $Y_1, \ldots, Y_p$ les variables mesurées sur chaque individu ;
- $\mathcal{Y}_1, \ldots, \mathcal{Y}_p$ les domaines d'observation associés à ces p variables.

L'objectif est de trouver une partition naturelle $P \in \mathcal{P}_k$ de l'ensemble E des individus en k classes où
- $P = \{C_1, \ldots, C_k\}$;
- $\mathcal{P}_k$ est l'ensemble des partitions de E en k classes.

Pour rechercher la meilleure partition, nous associons à chaque partition $P \in \mathcal{P}_k$ un critère de classification permettant d'évaluer sa qualité :

$$\begin{array}{rccc} W : & \mathcal{P}_k & \to & \mathbb{R} \\ & P & \rightsquigarrow & W(P,k). \end{array}$$

Le problème de la classification automatique consiste dès lors à rechercher la partition optimale $P^* = \{C_1^*, \ldots, C_k^*\}$ tel que

$$W(P^*,k) = \min_{P \in \mathcal{P}_k} W(P,k).$$

Un calcul élémentaire d'analyse combinatoire montre que le nombre de partitions possibles d'un ensemble de n individus croît plus qu'exponentiellement avec n. Il est donc impossible de chercher à optimiser le critère de classification sur toutes les partitions possibles.

Les méthodes de classification se restreignent à l'exécution d'un algorithme itératif convergent.

2.2 Structures classificatoires

Les méthodes de classification engendrent des structures classificatoires différentes. Nous définissons ici les partitions et les hiérarchies.

2.2.1 Les partitions

Une partition d'un ensemble E en k classes est un ensemble de parties non vides $P = \{C_1, \ldots, C_k\}$ d'intersections vides deux à deux et dont la réunion forme E, c'est-à-dire :

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

Exemple 2.2.1

Soit E un ensemble de 12 individus sur lesquels nous avons mesuré 2 variables. Graphiquement, une partition en 2 classes de ces individus est donnée par

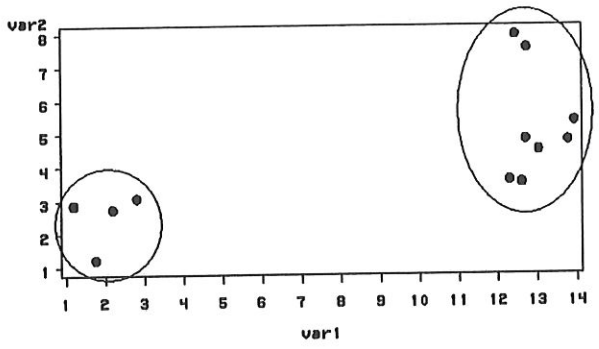

FIG. 2.1 – *Représentation graphique d'une partition.*

2.2.2 Les hiérarchies

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

Soient
- $E = \{x_1, \ldots, x_n\}$ un ensemble de n individus ;
- H un ensemble de parties (appelées paliers) non vides de E.

L'ensemble H est une hiérarchie sur E si et seulement si :
1. $E \in H$;
2. $\forall\, x_i \in E,\ \{x_i\} \in H$;
3. $\forall\, h,h' \in H,\ h \cap h' \neq \emptyset \ \Rightarrow\ h \subset h'$ ou $h' \subset h$.

Une hiérarchie de partitions est représentée graphiquement par un arbre hiérarchique, appelé dendrogramme.

Exemple 2.2.2

Soit $E = \{1,2,3,4,5\}$ un ensemble de 5 individus décrits par 2 variables.

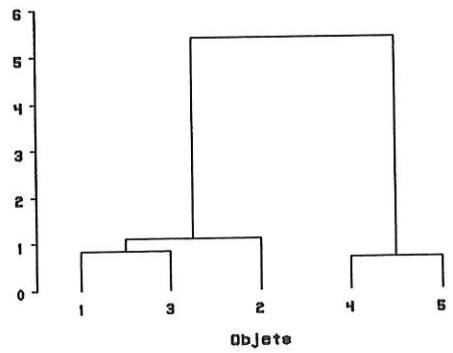

FIG. 2.2 – *Représentation graphique d'une hiérarchie.*

$$H = \Big\{\{1\},\{2\},\{3\},\{4\},\{5\},\{4,5\},\{1,3\},\{1,2,3\},\{1,2,3,4,5\}\Big\}.$$

2.3 Méthodes de classification

Il existe deux grandes familles de méthodes de classification :

- les méthodes hiérarchiques ;
- les méthodes de partitionnement.

2.3.1 Les méthodes hiérarchiques

Ces méthodes ont pour objectif de construire une suite de partitions de l'ensemble des individus en classes de moins en moins fines, de sorte que les regroupements successifs des classes forment une hiérarchie.

Deux types d'algorithmes peuvent être utilisés pour produire une telle suite de partitions :

- les algorithmes agglomératifs utilisés par les méthodes hiérarchiques ascendantes ;
- les algorithmes divisifs utilisés par les méthodes hiérarchiques descendantes.

2.3.1.1 Méthodes hiérarchiques ascendantes

Ces méthodes construisent une suite de partitions emboîtées en regroupant séquentiellement les classes deux à deux.

L'algorithme général de la classification hiérarchique ascendante est le suivant :

1. Partir de la partition discrète où chaque individu forme une classe ;
2. Regrouper, à chaque étape, les deux classes les plus proches au sens du critère d'agrégation choisi ;
3. Calculer la distance entre cette nouvelle classe et les autres ;
4. Répéter les pas 2 et 3 jusqu'à ce que les individus soient rassemblés au sein de la même classe.

Le choix du critère d'agrégation détermine la méthode de classification. Parmi les nombreuses méthodes de classification hiérarchiques ascendantes figurent les méthodes du lien simple et du lien complet, la méthode du centroïde et la méthode de Ward. Nous n'entrerons pas dans le détail en ce qui concerne ces méthodes.

2.3.1.2 Méthodes hiérarchiques descendantes

Les méthodes hiérarchiques descendantes, aussi appelées méthodes divisives fonctionnent dans le sens inverse. À chaque étape, une classe est divisée en deux jusqu'à obtenir n classes ne contenant plus qu'un seul individu.

La forme générale d'un algorithme divisif de classification est la suivante :

1. Partir de l'ensemble E des individus ;
2. Choisir, à chaque étape, la classe C à diviser ;
3. Déterminer $\{C_1,C_2\}$ une partition de cette classe C ;
4. Répéter les pas 2 et 3 jusqu'à l'obtention de la partition discrète où chaque individu forme une classe.

Les différents algorithmes divisifs de classification se distinguent par :

- le choix, à chaque étape, de la classe C à diviser ;
- le critère qui permet de scinder la classe C en deux classes C_1 et C_2.

2.3.2 Les méthodes de partitionnement

L'idée de base de ces méthodes est de choisir une partition initiale de l'ensemble des individus en un nombre de classes fixé a priori et de déplacer ces individus d'une classe à l'autre de façon itérative, de manière à améliorer la partition initiale.

La structure générale d'un algorithme itératif de partitionnement est la suivante :

1. *Initialisation :*
   - choisir une partition initiale de l'ensemble E en k classes,
   - déterminer le représentant initial de chaque classe ;
2. *Instructions itératives :*
   - calculer la dissimilarité entre chaque individu et chaque représentant de classe,
   - affecter les individus à la classe la plus proche,
   - déterminer le nouveau représentant de chaque classe ;
3. *Test d'arrêt :*
   - la partition reste inchangée ou
   - le nombre maximal d'itérations fixé par l'utilisateur est atteint.

Le choix de la partition initiale influence fortement le résultat final. C'est pourquoi, de manière pratique, nous appliquons plusieurs fois une méthode de partitionnement sur les données avec des initialisations différentes.

La méthode de partitionnement qui nous intéresse dans le cadre de ce mémoire est celle des nuées dynamiques. Nous présenterons cette méthode dans la section suivante.

2.4 Méthode des nuées dynamiques

2.4.1 La notion d'inertie

Soient
- $E = \{x_1, \ldots, x_n\}$ un ensemble d'individus sur lesquels nous avons mesuré p variables quantitatives ;
- un point $a \in \mathbb{R}^p$.

Les observations recueillies sont stockées dans la matrice

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

Nous appelons **inertie de E par rapport au point $a \in \mathbb{R}^p$** la quantité

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

où d est la distance euclidienne.

2.4.2 Théorème de Huygens

Huygens a montré 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 E par rapport à un point $a \in \mathbb{R}^p$ est liée à l'inertie par rapport au centre de gravité par la relation

$$I_a(E) = I_g(E) + n\, d^2(g, a).$$

L'inertie $I_g(E)$, notée T, est appelée inertie totale du nuage de points. Cette quantité mesure la dispersion des individus autour du centre de gravité g.

2.4.3 Inerties associées à une partition

Considérons
- $E = \{x_1, \ldots, x_n\}$ un ensemble d'individus décrits par p variables ;
- $P = \{C_1, \ldots, C_k\}$ 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 ;

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

À cette partition P, nous pouvons associer trois types d'inerties :

1. **L'inertie totale** T donnée par

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

L'inertie totale mesure la dispersion des individus autour du centre de gravité global g ;

2. **L'inertie intra-classe** W est définie par

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

L'inertie intra-classe est la somme sur chaque classe des distances entre les individus de la classe et leur centre de gravité. Cette quantité nous donne une indication sur l'homogénéité des classes ;

3. **L'inertie inter-classe** B est définie par

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

L'inertie inter-classe est l'inertie du nuage formé par les centres de gravité $g^{(l)}$ de chaque classe, pondéré par le nombre d'individus de la classe correspondante, par rapport au centre de gravité global g. Cette quantité mesure la dispersion des centres de gravité des classes autour du centre de gravité global g.

Les trois quantités que nous venons de définir sont reliées par la relation $T = W + B$.

L'inertie totale T est indépendante de la partition. Plus l'inertie inter-classe B est grande, plus l'inertie intra-classe W est petite. Minimiser l'inertie intra-classe revient donc à maximiser l'inertie inter-classe.

2.4.4 La méthode des nuées dynamiques

La méthode des nuées dynamiques a été introduite par E. Diday en 1971. Cette méthode cherche à optimiser un critère qui exprime l'adéquation entre une classification des individus et un mode de représentation des classes de cette classification. Le problème d'optimisation associé à cette méthode consiste donc à rechercher simultanément une classification des individus et une représentation des classes de cette classification parmi un ensemble de classifications et de représentations possibles, qui optimisent le critère.

2.4.4.1 Principe général

Nous associons tout d'abord à chaque classe d'individus un mode de représentation. Le prototype d'une classe peut être, par exemple, une droite, un groupe de points ou encore un centre du gravité.

Le déroulement de l'algorithme des nuées dynamiques est le suivant :

- k prototypes estimés ou tirés au hasard sont choisis parmi une famille de prototypes admissibles, appelée **espace de représentation**, notée $\mathcal{L}$;
- chaque individu est affecté au prototype le plus proche. Ainsi, nous obtenons une partition des individuss en k classes dont les k nouveaux prototypes sont calculés ;
- le procédé est alors recommencé avec les nouveaux prototypes.

Sous certaines conditions de régularité, cet algorithme fait décroître un critère W qui mesure l'adéquation entre les classes et leur prototype associé.

Le critère s'exprime de la manière suivante :

$$\begin{aligned} W : \ \mathcal{P}_k \times \mathcal{L}_k &\rightarrow \mathbb{R}^+ \\ (P,L) &\rightsquigarrow W(P,L) = \sum_{l=1}^{k} D(C_l,L^{(l)}) \end{aligned}$$

où

- $\mathcal{P}_k$ est l'ensemble des partitions P en k classes de E ;
- $L = \{L^{(1)},\ldots,L^{(k)}\} \in \mathcal{L}_k$ est un vecteur de k prototypes représentant les classes de P ;
- $D(C_l,L^{(l)})$ est une mesure d'adéquation du prototype $L^{(l)}$ à sa classe C_l. 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 leur prototype associé.

2.4.4.2 Plus formellement

L'objectif de la méthode 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 des classes

qui minimise un critère mathématique

$$\begin{array}{rccc} W : & \mathcal{P}_k \times \mathcal{L}_k & \rightarrow & \mathbb{R}^+ \\ & (P,L) & \rightsquigarrow & W(P,L) \end{array}$$

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

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 $W : \mathcal{P}_k \times \mathcal{L}_k \rightarrow \mathbb{R}^+$ qui permet de mesurer d'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.

Ce problème peut être résolu grâce à l'algorithme des nuées dynamiques. Celui-ci consiste à utiliser itérativement :

- une fonction de représentation $g : \mathcal{P}_k \rightarrow \mathcal{L}_k$;
- une fonction d'affectation $f : \mathcal{L}_k \rightarrow \mathcal{P}_k$.

La fonction de représentation g permet de calculer les k prototypes à partir de la partition en k classes, tandis que la fonction d'affectation f construit la partition en k classes en affectant chaque individu au prototype 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 $L^{(0)} \in \mathcal{L}_k$ estimée ou tirée au hasard.

2.4.4.3 Le cas du centre de gravité

Pour illustrer la méthode des nuées dynamiques, nous considérons le cas où les prototypes des classes sont les centres de gravité.

Espace de représentation

L'espace des individus et l'espace de représentation $\mathcal{L}$ d'une classe est l'espace $\mathbb{R}^p$. La mesure d'adéquation est définie par $D : \mathcal{P} \times \mathcal{L} \rightarrow \mathbb{R}^+$ tel que

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

où d est la 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

Par le théorème de Huygens, nous savons le centre de gravité est le point qui minimise l'inertie d'un nuage de points.

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

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

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

La fonction d'affectation f

La fonction d'affectation f est définie par

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

où

$$C_l = \left\{x \in \mathbb{R}^p \mid d(x,g^{(l)}) \leq d(x,g^{(m)})\ \forall\, m \in \{1,\ldots,k\}\right\}$$

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 $g^{(l)}$.

Le problème d'optimisation

Il s'agit de chercher le couple $(P^*,L^*) \in \mathcal{P}_k \times \mathcal{L}_k$ qui minimise 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

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

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

Nous pouvons également écrire ce critère sous la forme

$$W(P,L) = \sum_{l=1}^{k} W_l$$

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

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

Donc, dans le cas où les prototypes sont les k centres de gravité, la méthode des nuées dynamiques cherche à minimiser l'inertie intra-classe W et par conséquent à maximiser l'inertie inter-classe B en vertu de la relation $T = W + B$.

Chapitre 3

Méthodes de classification symboliques

Dans le cadre de ce mémoire, nous nous intéressons aux méthodes de classification symboliques incluses dans le logiciel SODAS 2 développé dans le cadre du projet ASSO. Le but de ce chapitre est de présenter et de décrire chacune de ces méthodes.

Nous commencerons notre présentation par la méthode de partitionnement SCLUST. Cette méthode est une extension symbolique de la méthode des nuées dynamiques classique [Diday, 1971]. Puis, nous décrirons la méthode DIV [Chavent, 1997] aussi appelée méthode de Chavent. Ensuite, nous parlerons de la méthode SCLASS [Rasson, Lallemand, 2000]. Pour terminer, nous présenterons la méthode HIPYR [Brito, 2000], laquelle permet de réaliser une classification hiérarchique ou pyramidale.

3.1 La méthode SCLUST

3.1.1 Présentation

La méthode de classification symbolique SCLUST est une extension de la méthode des nuées dynamiques classique [Diday, 1971] présentée dans le chapitre précédent.

Le but de cette méthode est de partitionner un ensemble de n objets symboliques décrits par un certain nombre de variables, en un nombre de classes k fixé a priori.

3.1.2 Principe général

La méthode SCLUST détermine de façon itérative une série de partitions qui améliore à chaque étape la valeur d'un critère de classification basé sur la proximité entre les objets et les prototypes qui représentent les classes.

Formellement, l'objectif de la méthode est de rechercher un couple $(P^*,L^*) \in \mathcal{P}_k \times \mathcal{L}_k$ où :

- $P^* \in \mathcal{P}_k$ est une partition optimale en k classes ;
- $L^* \in \mathcal{L}_k$ est un vecteur de k prototypes qui représentent les classes de P

qui minimise un critère d'ajustement entre P et L

$$\begin{array}{rccc} W: & \mathcal{P}_k \times \mathcal{L}_k & \rightarrow & \mathbb{R}^+ \\ & (P,L) & \rightsquigarrow & W(P,L) \end{array}$$

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

3.1.2.1 Prototypes pour les classes

Comme la méthode des nuées dynamiques, SCLUST nécessite la spécification de prototypes qui représentent les classes.

Considérons un ensemble $E = \{x_1,\ldots,x_n\}$ composé de n objets symboliques décrits par p variables $Y_1,\ldots,Y_p$.

- **Cas des variables intervalles**

  Chaque objet symbolique $x_i \in E$ est décrit par p variables intervalles notées Y_j $(j = 1,\ldots,p)$ et définies par

  $$\begin{array}{rccc} Y_j: & E & \rightarrow & \mathcal{B}_j \\ & x_i & \rightsquigarrow & Y_j(x_i) = x_{ij} = [\alpha_{ij},\beta_{ij}] \subset \mathbb{R}. \end{array}$$

  Chaque objet symbolique peut donc être représenté par un hyperrectangle dans un espace euclidien à p dimensions.

Les prototypes de chaque classe sont donc définis par les hyperrectangles de gravité respectifs.

L'hyperrectangle de gravité de la classe C_l est défini par :

$$L^{(l)} = \left(\left[\frac{1}{n_l} \sum_{x_i \in C_l} \alpha_{i1}, \frac{1}{n_l} \sum_{x_i \in C_l} \beta_{i1} \right], \ldots, \left[\frac{1}{n_l} \sum_{x_i \in C_l} \alpha_{ip}, \frac{1}{n_l} \sum_{x_i \in C_l} \beta_{ip} \right] \right)$$

où n_l désigne le nombre d'objets dans la classe C_l.

- **Cas des variables multivaluées**

  Chaque objet symbolique $x_i \in E$ est décrit par p variables multivaluées. La variable multivaluée Y_j dont l'espace d'observation est $\mathcal{Y}_j$ est définie par

$$\begin{array}{rccl} Y_j : & E & \rightarrow & \mathcal{Y}_j \\ & x_i & \rightsquigarrow & Y_j(x_i) = x_{ij} = \{c_1, \ldots, c_{m_j}\} \end{array}$$

  où m_j représente le nombre de catégories de la variable Y_j.

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

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

  La description symbolique d'un objet $x_i \in E$ est

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

  Le prototype de la classe C_l est défini par

$$L^{(l)} = \left(\frac{1}{n_l} \sum_{x_i \in C_l} (q_{1,x_i}(c_1), \ldots, q_{1,x_i}(c_{m_1})), \ldots, \frac{1}{n_l} \sum_{x_i \in C_l} (q_{p,x_i}(c_1), \ldots, q_{1,x_i}(c_{m_p})) \right)$$

  où n_l désigne le nombre d'objets dans la classe C_l.

- **Cas des variables modales**

  Le cas des variables modales est semblable à celui des variables multivaluées. Les fréquences $q_{j,x_i}(c_s)$ sont simplement remplacées par les valeurs de la distribution π_{j,x_i} associées à chacune des catégories de $Y_j(x_i)$.

3.1.2.2 Espace de représentation

Dans le cas de variables intervalles, l'espace de représentation des objets symboliques est 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

$$D : \mathcal{P} \times \mathcal{L} \rightarrow \mathbb{R}^+$$

tel que

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

où

- d est une mesure de dissimilarité ;
- $\mathcal{P}$ est l'ensemble des parties de E.

Les trois types de distances définis dans le chapitre 1, à savoir les distances $\mathcal{L}^1$, $\mathcal{L}^2$ et la distance de Hausdorff sont utilisables dans SCLUST.

3.1.2.3 Fonction de représentation

La fonction de représentation g qui, à toute partition $P = (C_1,\ldots,C_k)$ en k classes associe sa représentation $L = (L^{(1)},\ldots,L^{(k)})$, est définie par

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

où $L^{(l)}$ est le prototype de la classe C_l.

3.1.2.4 Fonction d'affectation

La fonction d'affectation f est définie par

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

où

$$C_l = \left\{x \in E \mid d\left(x,L^{(l)}\right) \leq d\left(x,L^{(m)}\right), \forall\, m \in \{1,\ldots,k\}\right\}$$

c'est-à-dire que C_l est la classe composée des objets les plus proches, au sens de la métrique choisie, de son prototype $L^{(l)}$.

3.1.2.5 Le problème d'optimisation

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

$$W(P,L) = \sum_{l=1}^{k} D(C_l,L^{(l)})$$

où $D(C_l,L^{(l)})$ est la mesure d'adéquation de la classe C_l à son représentant $L^{(l)}$.

Une décroissance de W exprime une meilleure adéquation entre les classes et les prototypes associés.

Nous pouvons réécrire le critère sous la forme

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

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

La méthode consiste à minimiser l'inertie intra-classe W et donc à maximiser l'inertie inter-classe B en vertu de la relation $T = W + B$.

3.1.3 Algorithme

1. *Initialisation*
   Déterminer une partition aléatoire $P^{(0)}$.
2. *Etape de représentation*
   Pour $l = 1,\ldots,k$ calculer le prototype $L^{(l)}$ associé à la classe C_l.
3. *Etape d'affectation*
   Un à un les objets de E sont affectés à la classe dont le prototypes est le plus proche.
4. *Test d'arrêt*
   Si la partition est stable ou **si** le nombre maximal d'itérations fixé est atteint
   $\longrightarrow$ STOP
   Sinon répéter les pas 2 et 3.

3.2 La méthode DIV

3.2.1 Présentation

La méthode DIV [Chavent, 1997] est une méthode de classification symbolique hiérarchique monothétique divisive. Elle construit une hiérarchie de partitions à partir de l'ensemble des objets symboliques, en divisant successivement une classe en deux sous-classes, et ce en ne tenant compte que d'une seule variable à la fois.

À chaque étape de l'algorithme, une classe est sélectionnée pour être divisée selon une question binaire. Cette question induit la meilleure bipartition conformément à un critère d'évaluation.

Dans ce qui suit, nous allons voir comment est déterminée cette meilleure bipartition et comment la classe à diviser est choisie par DIV.

3.2.2 Extension du critère de la variance intra-classe

Le critère d'évaluation utilisé par Chavent est une généralisation de la notion d'inertie intra-classe au cas des données symboliques.

L'inertie intra-classe est un critère classique d'évaluation d'une partition. Il permet notamment de mesurer l'homogénéité des classes de la partition.

L'inertie intra-classe de la classe C_l est donnée par

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

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

Montrons que

$$I(C_l) = \sum_{x_i \in C_l} d^2(x_i, g^{(l)}) = \frac{1}{2\,n_l} \sum_{x_i \in C_l} \sum_{x_j \in C_l} d^2(x_i, x_j). \tag{3.1}$$

Preuve :

Par le théorème de Huygens, nous savons que l'inertie de la classe C_l par rapport à un objet $x_i \in C_l$ est donnée par

$$I_{x_i}(C_l) = I_{g^{(l)}}(C_l) + n_l\, d^2(x_i, g^{(l)}) \tag{3.2}$$

où n_l désigne le nombre d'objets dans la classe C_l [1].

1. Nous supposons tous les poids égaux à 1.

En sommant sur tous les objets $x_i \in C_l$, nous obtenons

$$\sum_{x_i \in C_l} I_{x_i}(C_l) = n_l\, I_{g^{(l)}}(C_l) + n_l \underbrace{\sum_{x_i \in C_l} d^2(x_i, g^{(l)})}_{I_{g^{(l)}}(C_l)} = 2\, n_l\, I_{g^{(l)}}(C_l). \tag{3.3}$$

Par définition, l'inertie de la classe C_l par rapport à un objet $x_i \in C_l$ est donnée par

$$I_{x_i}(C_l) = \sum_{x_j \in C_l} d^2(x_i, x_j). \tag{3.4}$$

En injectant l'expression (3.4) dans (3.3), nous obtenons

$$\sum_{x_i \in C_l} \sum_{x_j \in C_l} d^2(x_i, x_j) = 2\, n_l\, I_{g^{(l)}}(C_l). \tag{3.5}$$

D'où le résultat

$$I_{g^{(l)}}(C_l) = \frac{1}{2\, n_l} \sum_{x_i \in C_l} \sum_{x_j \in C_l} d^2(x_i, x_j) \tag{3.6}$$

où d est une dissimilarité entre objets symboliques.

□

Ainsi, la quantité

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

est l'inertie intra-classe définie sur un tableau de dissimilarités.

L'algorithme recherche donc la partition de sorte que l'inertie intra-classe soit minimale.

3.2.3 Bipartition d'une classe

Le problème consiste à trouver une partition en deux classes $\{C_l^1, C_l^2\}$ de la classe C_l de manière à ce que la partition obtenue soit optimale au sens du critère d'évaluation choisi. Ici, l'algorithme recherche la bipartition qui minimise l'inertie intra-classe.

Pour trouver cette partition optimale en deux classes, Chavent propose de rechercher la meilleure bipartition parmi toutes les bipartitions induites par un ensemble de questions binaires.

3.2.3.1 Question binaire

Une question binaire est une condition à laquelle les objets satisfont ou ne satisfont pas

$$Y_j \in V \quad \text{ou} \quad Y_j \in \overline{V} \ ?$$

où $\{V,\overline{V}\}$ est une partition en deux classes du domaine d'observation $\mathcal{Y}_j$ associé à la variable Y_j.

Lorsque le domaine d'observation $\mathcal{Y}_j$ est ordonné, la partition $\{V,\overline{V}\}$ est entièrement définie par une valeur $c \in \mathcal{Y}_j$, appelée valeur de coupure. La question binaire est alors souvent représentée par

$$Y_j \leq c \quad \text{ou} \quad Y_j > c \ ?$$

À chaque question binaire est alors associée une fonction binaire

$$q_c : E \rightarrow \{0,1\}$$

tel que

$$q_c(x_i) = \begin{cases} 0 & \text{si } Y_j(x_i) \leq c \\ 1 & \text{sinon.} \end{cases}$$

La bipartition $\{C_l^1,C_l^2\}$ de la classe C_l induite par la question binaire q_c est définie par

$$C_l^1 = \{x_i \in C_l \mid q_c(x_i) = 0\}$$
$$C_l^2 = \{x_i \in C_l \mid q_c(x_i) = 1\}.$$

3.2.3.2 Détermination de la valeur de coupure et choix de la meilleure bipartition

L'objectif de cette section est de déterminer de manière constructive une valeur de coupure c et de trouver ensuite la meilleure bipartition.

Dans le cas des variables intervalles, la fonction q_c associée à la question "$Y_j \leq c$" est définie par

$$q_c(x_i) = \begin{cases} 0 & \text{si } m_{x_i} \leq c \\ 1 & \text{si } m_{x_i} > c \end{cases}$$

où la quantité $m_{x_i} = \frac{\alpha+\beta}{2}$ est appelée valeur médiane de la description de l'objet x_i sur la variable Y_j.

Soient

- z_j le nombre de bipartitions induites par les questions binaires sur la variable Y_j ;
- n le nombre d'objets de la classe C considérée.

Quelle que soit la valeur de coupure c entre deux valeurs médianes consécutives m_k et m_{k+1}, la partition induite est la même. Nous avons donc au plus $z_j = n - 1$ bipartitions induites par les questions binaires sur la variable Y_j.

Pour ne poser que $n-1$ questions pour générer toutes ces bipartitions, Chavent décide d'utiliser les $n-1$ valeurs de coupure définies par

$$c = \frac{m_k + m_{k+1}}{2} \quad (k = 1 \ldots, n-1)$$

c'est-à-dire les centres des intervalles $[m_k, m_{k+1}]$.

Pour chacune des $n-1$ valeurs de coupure, l'algorithme calcule la perte d'inertie intra-classe occasionnée par la division de la classe C en deux sous-classes C^1 et C^2 donnée par

$$\triangle C = I(C) - I(C^1) - I(C^2),$$

où $I(C) = \frac{1}{2n} \sum_{x_i \in C} \sum_{x_j \in C} d^2(x_i, xj)$.

Nous retenons la partition $\{C^1, C^2\}$ pour laquelle la quantité $\triangle C$ est maximale, ainsi que la valeur de coupure correspondante.

Cette partition est celle pour laquelle la division de la classe C occasionne les deux sous-classes

- les plus homogènes possible ;
- les plus distinctes possible.

3.2.3.3 Détermination de la meilleure variable

Pour chaque variable Y_j mesurée sur les objets, l'algorithme retient la meilleure bipartition au sens de l'inertie.

Si z_j est le nombre de bipartitions induites par la variable Y_j, l'algorithme sélectionne, parmi les $z_1 + z_2 + \cdots + z_p$ bipartitions de la classe C, celle pour laquelle l'inertie intra-classe est minimale.

3.2.4 Choix de la classe à diviser

Soit $P^{(k)} = \{C_1,C_2,\ldots,C_k\}$ la partition de l'ensemble des objets obtenue à l'étape k.

Il faut à présent choisir une classe $C_j \in P^{(k)}$ qui sera scindée en deux sous-classes C_j^1 et C_j^2 . Dans sa méthode, Chavent propose de choisir la classe $C_j \in P^{(k)}$ à diviser de telle sorte que la nouvelle partition

$$P^{(k+1)} = P^{(k)} \cup \{C_j^1,C_j^2\} - \{C_j\}$$

ait une inertie intra-classe

$$W(P^{(k+1)}) = W(P^{(k)}) - I(C_j) + I(C_j^1) + I(C_j^2)$$

minimale.

Autrement dit, le problème consiste à déterminer la classe $C_j \in P^{(k)}$ qui minimise l'inertie intra-classe $W(P^{(k+1)})$, c'est-à-dire celle qui maximise la perte d'inertie

$$\triangle C_j = I(C_j) - I(C_j^1) - I(C_j^2).$$

3.2.5 Algorithme

1. *Initialisation*
   Partir de la partition initiale $P^{(0)} = E$ où E désigne l'ensemble des objets.
2. *Choix de la meilleure bipartition*
   Pour chaque classe C de la partition, choisir dans l'ensemble des partitions induites par l'ensemble des questions binaires sur C celle qui minimise

$$W = \frac{1}{2\,n_1} \sum_{x_i \in C_1} \sum_{x_j \in C_1} d^2(x_i,xj) + \frac{1}{2\,n_2} \sum_{x_i \in C_2} \sum_{x_j \in C_2} d^2(x_i,xj)$$

3. *Choix de la classe à diviser*
   Choisir la classe C de la partition tel que la perte d'inertie

$$\triangle C = I(C) - I(C^1) - I(C^2)$$

   est maximale.
4. *Mise à jour de la partition*
5. *Test d'arrêt*
   Si le nombre maximal d'itérations fixé par l'utilisateur est atteint
   $\longrightarrow$ STOP
   Sinon répéter les pas 2, 3 et 4.

3.2.6 Résultats de l'algorithme

En sortie, nous obtenons une hiérarchie de partitions. Les singletons de cette hiérarchie sont les classes de la partition obtenue à la dernière étape de l'algorithme.

Chaque classe formée peut être associée à une conjonction de propriétés exprimées en termes des variables initiales, c'est-à-dire à un objet symbolique assertion

$$q = \bigwedge_{m=1}^{r} [Y_{j_m} \mathcal{R}_{j_m} z_m]$$

où z_m est un point de l'espace d'observation $\mathcal{Y}_j$.

Les classes formées peuvent donc être considérées comme des concepts décrits en extension par l'ensemble des objets qui la composent et en intension par un objet symbolique assertion qui exprime ses propriétés. Nous obtenons ainsi une première description des classes.

3.3 La méthode SCLASS

La méthode des arbres de clustering, ou méthode SCLASS est une méthode de classification symbolique monothétique divisive. Elle est une extension de la méthode classique UNHOPPKI proposée par Pirçon dans sa thèse.

La méthode SCLASS s'applique uniquement aux données de type intervalle. Elle a pour objectif de diviser successivement les noeuds en choisissant la meilleure variable intervalle.

L'originalité de cette méthode réside dans la manière dont est coupé un noeud. En effet, la coupure d'un noeud se base sur l'hypothèse que les distributions des points peuvent être modélisées par des processus de Poisson non-homogènes dont l'intensité sera estimée par la méthode des noyaux. La coupure se fera entre les modes de cette densité de façon à maximiser la fonction de vraisemblance.

3.3.1 Le processus de Poisson non-homogène

Le **P**rocessus de **P**oisson **N**on-**H**omogène (PPNH) d'intensité $q(\cdot)$ sur le domaine $D \in R^p$ est caractérisé par les deux propriétés suivantes :

1. $\forall A \subset D, N(A)$ a une distribution de Poisson de paramètre $\int_A q(x)\, m(dx)$, où m est la mesure de Lebesgue ;
2. Si $N(A) = n$, alors les n points sont distribués indépendamment dans A, avec une fonction de densité proportionnelle à $q(x)$.

3.3.2 Hypothèse générale

La méthode de classification symbolique SCLASS est basée sur une seule hypothèse: les points observés sont générés par un processus de Poisson non-homogène N d'intensité $q(\cdot)$ dans $D \in R^p$, où D est l'union de k domaines disjoints convexes $D_1,\ldots,D_k$.

La fonction de vraisemblance, pour les observations $\underline{x} = (x_1,x_2,\ldots,x_n)$ avec $x_i \in \mathbb{R}^p$, $(i = 1,\ldots,n)$ vaut

$$L_D(\underline{x}) = \frac{1}{(\rho(D))^n} \prod_{i=1}^{n} \mathbb{1}_D(x_i) \cdot q(x_i)$$

où

- $q(\cdot)$ est l'intensité du processus;
- $\rho(D) = \int_D q(x)\,dx$ est l'intensité intégrée du processus;
- $\mathbb{1}_D(.)$ est la fonction indicatrice.

Si l'intensité du processus est connue, la solution du maximum de vraisemblance correspondra aux k domaines convexes disjoints contenant tous les points et pour lesquels la somme de leurs intensités intégrées est minimale. Lorsque l'intensité du processus est inconnue, il faut l'estimer.

3.3.3 La méthode des noyaux

Pour estimer l'intensité du processus de Poisson non-homogène, une méthode d'estimation de densité non-paramétrique est utilisée: la méthode des noyaux.

Cet estimateur de densité, appelé estimateur noyau, est défini par

$$\hat{q}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

où

- h est la largeur de la fenêtre (paramètre de lissage);
- K est une fonction de poids positive, symétrique telle que $\int_{-\infty}^{+\infty} K(x)\,dx = 1$.

L'estimateur noyau est une somme de "bosses" centrées sur les observations. La fonction de poids K, encore appelée fonction noyau, détermine la forme des bosses tandis que le paramètre h détermine leur largeur.

Le choix du paramètre de lissage h est une étape essentielle de l'estimation dans la mesure où

- si le paramètre de lissage h est trop petit, le lissage est insuffisant : l'estimateur dégénère en une suite de n pics localisés sur les points de l'échantillon ;
- si le paramètre de lissage h est trop grand, le lissage est trop important : l'estimation se rapproche de celle d'une loi uniforme et cela engendre une perte d'information.

3.3.4 Recherche de "bosses" et test de la multi-modalité

Silverman distingue les notions de bosses et de modes par les définitions suivantes : un mode dans une densité f sera un maximum local, tandis qu'une bosse sera caractérisée par un intervalle $[a,b]$ de telle façon que la densité f soit concave sur cet intervalle mais pas sur un intervalle plus grand.

Dans le cadre de l'estimation de la densité par la méthode des noyaux, le nombre de modes est déterminé par le paramètre de lissage h. Pour de très grandes valeurs de h, l'estimation $\hat{q}$ de la densité sera unimodale. Par contre au fur et à mesure que h diminue, le nombre de modes augmente.

Ce comportement est décrit mathématiquement comme *"le nombre de modes est une fonction décroissante de la largeur de la fenêtre h"*. Ceci est garanti seulement pour certains noyaux, tel le noyau normal.

Par conséquent, pour estimer l'intensité du processus de Poisson non-homogène, c'est la méthode du noyau qui est utilisée avec le noyau normal défini par

$$K_{\mathcal{N}}(y) = \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}}.$$

Étant donné qu'un noyau normal est utilisé, il existe une valeur critique h_{crit} du paramètre de lissage pour lequel l'estimation change de l'unimodalité à la multi-modalité. C'est cette valeur critique que recherche le critère de coupure.

3.3.5 La règle de coupure

Le problème ici consiste en la détermination de la coupure optimale au sens du maximum de vraisemblance.

L'algorithme de SCLASS choisit la valeur h pour laquelle l'estimation de la densité est multi-modale tout en ayant le plus petit nombre de modes.

Une fois ce paramètre h déterminé, l'algorithme scinde le domaine D en deux domaines

D_1 et D_2 disjoints convexes pour lesquels la fonction de vraisemblance

$$L_{D_1,D_2}(\underline{x}) = \frac{1}{\left(\rho(D_1) + (\rho(D_2)\right)^n} \prod_{i=1}^{n} \mathbb{1}_{D_1 \cup D_2}(x_i) \cdot \hat{\mathrm{q}}(x_i)$$

est maximale, c'est-à-dire pour lesquels l'aire intégrée

$$\rho(D_1) + \rho(D_2)$$

est la plus petite.

En procédant de la sorte variable par variable, l'algorithme parvient à déterminer celle qui engendre la plus grande fonction de vraisemblance.

L'algorithme s'arrête lorsqu'un nombre minimal d'effectif par noeud est atteint.

3.3.6 L'élagage

A l'issue du processus de coupure, nous obtenons un arbre de grande taille que nous souhaitons simplifier. C'est pourquoi une méthode d'élagage a été mise en place.

La procédure d'élagage construite fait également l'hypothèse que les points sont la réalisation d'un processus de Poisson non-homogène. Elle est basée sur un test d'hypothèse appelé le *Gap test.*

Dans le cas de deux classes D_1 et D_2 ($D_1 \cup D_2 = D$), le *Gap test* consiste à tester l'hypothèse nulle

$$H_0 : \text{il y a } n = n_1 + n_2 \text{ points dans } D_1 \cup D_2$$

contre l'hypothèse alternative

$$H_1 : \text{il y a } n_1 \text{ points dans } D_1 \text{ et } n_2 \text{ points dans } D_2 \text{ avec } D_1 \cap D_2 = \emptyset.$$

Autrement dit, l'hypothèse nulle signale que le noeud est terminal, contrairement à l'hypothèse alternative qui justifie la coupure du noeud en ses deux fils. L'hypothèse nulle entraîne ce que nous appelons une mauvaise coupure alors que l'hypothèse alternative donne une bonne coupure.

En parcourant l'arbre branche par branche, de la racine vers les feuilles il est possible de détecter les bonnes et les mauvaises coupures. Les fins des branches pour lesquelles il n'y a que des mauvaises coupures sont élaguées.

3.3.7 Application aux données de type intervalle

Considérons un ensemble $E = \{x_1,\ldots,x_n\}$ composé de n objets symboliques caractérisés par p variables intervalles.

La variable Y_j est définie par

$$\begin{array}{rcl} Y_j : E & \to & \mathcal{B}_j \\ x_k & \rightsquigarrow & Y_j(x_i) = x_{ij} = [\alpha_{ij},\beta_{ij}] \subset \mathbb{R}. \end{array}$$

Chaque intervalle est représenté par ses coordonnées Milieu-Longueur dans l'espace $(M,L) \subset \mathbb{R} \times \mathbb{R}^+$.

Dans l'espace (M,L), les intervalles sont représentés par des points. Nous considérons toutes les partitions en deux classes en effectuant des coupures parallèles à l'axe des longueurs.

La méthode SCLASS étant une méthode divisive, les coupures doivent respecter l'ordre des centres des classes.

Nous recherchons l'intervalle $]M_i,M_{i+1}[$ tel que l'intensité intégrée

$$\int_{M_i}^{M_{i+1}} q_1(x)\, dm(x) + \int_{\min(L_i,L_{i+1})}^{\max(L_i,L_{i+1})} q_2(y)\, dm(y) \tag{3.7}$$

où

- q_1 est l'intensité sur l'axe M;
- q_2 est l'intensité sur l'axe L

est maximale.

Nous choisissons la bipartition générée par n'importe quelle valeur situé dans l'intervalle $]M_i,M_{i+1}[$ qui maximise (3.7).

Un noeud C est divisé en deux sur base de la réponse à une question binaire de type "$m_{ij} \leq c$" où c est la valeur de coupure.

À cette question binaire, nous associons une fonction binaire $q_c : E \to \{0,1\}$ tel que

$$q_c(x_i) = \begin{cases} 0 & \text{si } m_{ij} \leq c \\ 1 & \text{sinon.} \end{cases}$$

La bipartition du noeud C en C_1 et C_2 est donnée par

- $C_1 = \{x \in C \mid q_c(x) = 0\}$
- $C_2 = \{x \in C \mid q_c(x) = 1\}$.

3.3.8 Algorithme

1. *Initialisation*
   Partir de la partition initiale $P^{(0)} = E$ où E désigne l'ensemble des objets.
2. *Choix de la meilleure bipartition*
   Pour chaque noeud (classe) et pour chaque variable
   - Estimer l'intensité $q(x)$,
   - Trouver la meilleure séparation de C en C_1 et C_2 tel que $C = C_1 \cup C_2$ et $\rho(C_1) + \rho(C_2)$ est minimal, c'est-à-dire tel que la perte intégrée de l'intensité est maximale.
3. *Choix de la classe à diviser*
   Choisir le noeud et la variable $V_{\max}$ qui donnent la fonction de vraisemblance maximale.
4. *Règle de coupure*
   Couper la classe pour laquelle $\rho(C_1) + \rho(C_2)$ est minimal.
5. *Mise à jour de la partition*
6. *Test d'arrêt*
   Si l'effectif minimal par noeud est atteint
   $\longrightarrow$ STOP
   Sinon répéter les pas 2, 3, 4 et 5.
7. *Elagage*
   Passer en revue les noeuds de l'arbre et élaguer les fins de branches pour lesquelles il n'y a que des mauvaises coupures.

3.3.9 Résultats de l'algorithme

En sortie, nous obtenons un arbre de classification. Les noeuds de l'arbre représentent les questions binaires sélectionnées par l'algorithme. Les k feuilles définissent la partition de l'ensemble des objets en k classes.

Comme pour la méthode DIV, les classes peuvent être associées chacune à un objet symbolique assertion exprimant leurs propriétés.

3.4 La méthode HIPYR

3.4.1 Présentation

La méthode de classification symbolique HIPYR permet de réaliser une classification ascendante hiérarchique ou pyramidale sur un ensemble d'objets décrits par un certain nombre de variables symboliques.

Le but de cette méthode est de regrouper successivement les objets symboliques de manière obtenir des classes homogènes.

3.4.2 Principe général

Le principe de la méthode HIPYR est un peu particulier dans le sens où la hiérarchie de partitions peut être construite de deux façons différentes :

- soit à partir de la matrice de données symboliques ;
- soit à partir de la matrice des dissimilarités entre objets symboliques.

Nous verrons dans ce qui suit que les deux manières de procéder sont totalement différentes.

3.4.2.1 Classification des objets à partir de la matrice de données

Dans ce cas, les classes sont considérées comme des concepts. Chaque classe est associée à un objet symbolique de type assertion, c'est-à-dire un objet symbolique qui généralise les objets membres de la classe de manière à ce qu'aucun objet en dehors de la classe ne corresponde à la description donnée par l'objet symbolique assertion.

Une classe peut donc être représentée par un couple (C,s) où

- C est l'ensemble des objets dans la classe ;
- s est l'objet symbolique assertion contenant l'extension de C.

L'algorithme de classification est un algorithme ascendant. Il débute donc avec pour classes les n singletons. À chaque étape de l'algorithme, un nouveau couple (C,s) est formé par la réunion de deux couples (C_1,s_1) et (C_2,s_2) déjà construits.

Les classes C_1 et C_2 sont sélectionnées pour être réunies si les deux conditions suivantes sont satisfaites :

1. les classes C_1 et C_2 peuvent effectivement être agrégées, c'est-à-dire dans le cas hiérarchique qu'aucune des deux classes n'a déjà été agrégée dans un des pas précédents ;
2. l'objet symbolique assertion $s = s_1 \cup s_2$ généralise les objets membres de la classe $C = C_1 \cup C_2$ de manière à ce qu'aucun objet en dehors de C ne corresponde à la description donnée par s.

La construction de $s = s_1 \cup s_2$ est appelée étape de généralisation.

Plusieurs classes peuvent remplir ces conditions ce qui nécessite la définition d'un critère évaluation permettant de choisir la meilleure agrégation parmi toutes celles possibles.

Le critère d'évaluation utilisé par la méthode HIPYR est le critère du degré de généralité. Dans le cas de variables intervalles ou multivaluées catégoriques, ce critère mesure la proportion du domaine d'observation couverte par l'objet symbolique assertion associé à la classe. Dans le cas de variables modales, ce critère évalue dans quelle mesure la distribution donnée est proche d'une distribution uniforme.

Parmi toute les paires $(C_1,s_1),(C_2,s_2)$ qui satisfont les conditions énoncées précédemment, l'algorithme réunit celles pour lesquelles le couple résultant (C,s) a la plus petite valeur du critère.

Si à un moment donné, il n'y a pas de paires de classes qui satisfont les conditions d'agrégation, l'algorithme recherche si le regroupement de plusieurs classes est possible. Dans ce cas, c'est l'ensemble de toutes ces classes qui doit vérifier les conditions d'agrégation.

3.4.2.2 Classification des objets à partir de la matrice de dissimilarités

Dans ce cas, la classification est basée sur une matrice de dissimilarités entre objets symboliques. Cette matrice de dissimilarités doit obligatoirement être calculée à l'aide du module DISS inclus dans le logiciel SODAS.

À chaque étape de l'algorithme, les objets symboliques les plus semblables sont regroupés ensemble. Les méthodes de classification hiérarchiques proposées par HIPYR sont les suivantes :

- *la méthode du lien simple*
  La distance entre deux classes C_i et C_j est définie par la plus petite distance séparant

un objet de la classe C_i et un objet de la classe C_j :

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

- *la méthode du lien complet*
  La distance entre deux classes C_i et C_j est définie par la plus grande distance entre un objet de la classe C_i et un objet la classe de C_j :

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

- *la méthode de la distance moyenne*
  Ce critère consiste à calculer la distance moyenne entre tous les objets de C_i et tous les objets de C_j :

$$d(C_i,C_j) = \frac{1}{n_i \cdot n_j} \sum_{x \in C_i} \sum_{y \in C_j} d(x,y),$$

  où n_i et n_j désignent respectivement le nombre d'objets dans les classes C_i et C_j;
- *la méthode du diamètre*

3.4.3 Algorithme

1. *Initialisation*
   Chaque objet symbolique forme une classe.
2. *Etape d'agrégation*
   - Pour une classification à partir de la matrice des données :
     - sélectionner les paires (C_i,s_i) et (C_j,s_j) qui vérifient les conditions d'agrégation,
     - choisir la meilleure paire parmi l'ensemble de paires possibles ;
   - Pour une classification à partir de la matrice des dissimilarités :
     - regrouper les deux classes les plus proches au sens du critère d'agrégation choisi,
     - calculer la distance entre cette nouvelle classe et les autres.
3. *Test d'arrêt*
   Si tous les objets de l'ensemble de départ sont réunis au sein d'une même classe $\longrightarrow$ STOP.
   Sinon retourner à l'étape d'agrégation.

Chapitre 4

Comparaison des méthodes sur des données artificielles

Dans ce chapitre, nous réalisons une étude comparative des méthodes de classification symboliques présentées dans le chapitre 3.

Nous allons tester les méthodes de classification symboliques sur plusieurs jeux de données artificielles, chacun présentant une structure de données différentes :

- données avec deux classes hypersphériques séparées ;
- données avec trois classes hypersphériques séparées ;
- données avec deux classes allongées ;
- données avec deux classes emboîtées.

Nous analyserons les résultats obtenus par chacune des méthodes sur les quatre jeux de données. Finalement, nous dresserons le bilan de notre étude.

Nous travaillerons uniquement avec des objets symboliques décrits par des variables intervalles. En effet, il s'agit du seul type de données symboliques qui est traité par toutes les méthodes de classification envisagées.

4.1 Données avec deux classes hypersphériques

4.1.1 Le jeu de données

Nous débutons cette étude par l'analyse d'un jeu de données composé de 10 objets symboliques décrits par 2 variables intervalles.

Pour créer ce jeu de données, nous avons généré les minima et maxima de chaque intervalle suivant chacune des deux variables de manière à obtenir deux classes hypersphériques bien séparées, à l'aide d'un outil de génération de nombres aléatoires en Excel.

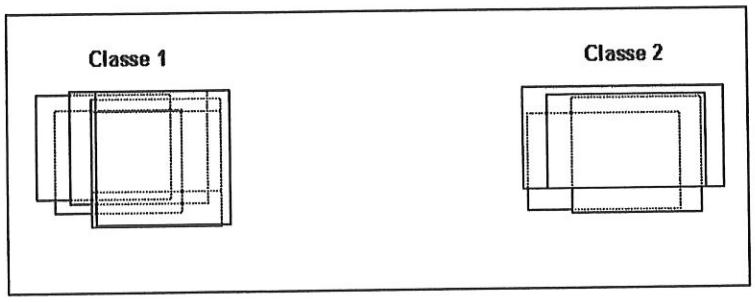

FIG. 4.1 – *Représentation graphique des objets symboliques.*

Pour clarifier les sorties du logiciel SODAS, nous avons labélisé les objets symboliques comme suit :

- classe 1 : `S01_x` ;
- classe 2 : `S02_y`.

4.1.2 La méthode SCLUST

Pour ce qui est de l'initialisation de l'algorithme, nous avons opté pour une initialisation par prototypes[1]. Lors d'une initialisation par prototypes, un nombre réel strictement compris entre zéro et le nombre d'objets symboliques contenus dans le jeu de données étudié, est associé aléatoirement à chaque classe. Ce nombre réel est ensuite tronqué à l'entier supérieur. C'est ce nombre tronqué qui détermine l'objet symbolique qui servira de représentant de la classe considérée.

Étant donné que la partition finale obtenue à l'issue de cette méthode dépend fortement de l'initialisation[2], nous avons décidé de réinitialiser 50 fois l'algorithme. Le fait de réinitialiser plusieurs fois l'algorithme permet d'optimiser le critère de l'inertie et donc d'améliorer la partition finale.

1. Option *Random prototypes*.
2. Pour rappel, la méthode SCLUST est une méthode de partitionnement.

Nous avons testé la méthode SCLUST avec les trois distances définies pour les données de type intervalle, à savoir la distance de Hausdorff, la distance $\mathcal{L}^1$ et la distance $\mathcal{L}^2$.

La partition optimale en deux classes trouvée par SCLUST avec chacune des trois distances est la suivante :

```
Classe :   1  Cardinal :    6
==============================

( 0) SO1_1    [2.3]  ( 1) SO1_2    [0.8]  ( 2) SO1_3    [0.1]
( 3) SO1_4    [0.7]  ( 4) SO1_5    [0.8]  ( 5) SO1_6    [1.2]

Classe :   2  Cardinal :    4
==============================

( 6) SO2_1    [1.3]  ( 7) SO2_2    [1.3]  ( 8) SO2_3    [0.2]
( 9) SO2_4    [1.2]
```

Comme nous pouvons le constater, cette partition en deux classes correspond bien à la partition naturelle des données.

Les prototypes des classes[3] pour chacune des deux variables sont représentés à la figure 4.2 :

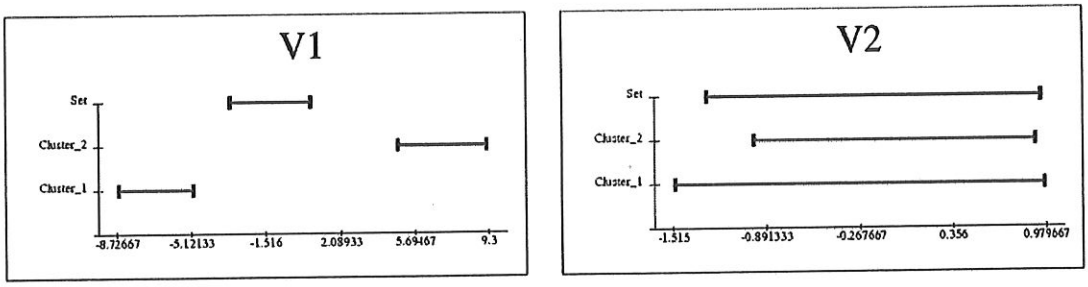

FIG. 4.2 – *Prototypes des deux classes pour les variables* V1 *et* V2.

Pour la variable V1, les prototypes des classes sont bien séparés. Ceci souligne le pouvoir discriminant de cette variable.

3. Avec la distance $\mathcal{L}^1$.

La description de cette partition est reprise dans le tableau 4.1 :

Distances	Nombre de fois où la partition naturelle en 2 classes est retrouvée	Valeur initiale du critère	Valeur finale du critère	Pourcentage d'inertie expliquée par la partition
Hausdorff	41	62.41	9.45	84.86
$\mathcal{L}^1$	50	138.53	15.52	88.79
$\mathcal{L}^2$	50	930.62	8.65	99.07

TAB. 4.1 – *Résultats de la méthode* SCLUST.

Au vu de ces résultats, nous pouvons dire que la méthode SCLUST est efficace pour cet exemple. Nous pouvons constater que la partition naturelle des données en deux classes est retrouvée pratiquement à chaque fois à la fin de l'algorithme.

Par ailleurs, en assimilant les objets symboliques aux prototypes des classes les plus proches, le pourcentage d'inertie expliqué par la partition est très élevé. Il est de 85% pour la distance de Hausdorff, de 89% pour la distance $\mathcal{L}^1$ et de 99% pour la distance $\mathcal{L}^2$.

4.1.3 La méthode DIV

La méthode DIV a fournit la hiérarchie de partitions suivante :

```
PARTITION IN 2 CLUSTERS  :
-------------------------

Cluster 1 (n=6) : SO1_1 SO1_2 SO1_3 SO1_4 SO1_5 SO1_6
Cluster 2 (n=4) : SO2_1 SO2_2 SO2_3 SO2_4

Explicated inertia : 98.831777
```

Cette partition en deux classes correspond bien à la partition naturelle des données.

La description de la partition trouvée est la suivante :

```
Cluster 1 :
   IF    1- [V1 <= 0.167500] IS TRUE

Cluster 2 :
   IF    1- [V1 <= 0.167500] IS FALSE
```

Comme nous pouvons le voir, la coupure de l'ensemble des objets symboliques se fait suivant la variable V1, ce qui souligne une fois de plus le pouvoir discriminant de cette variable.

Cette classification des objets symboliques en deux classes fait apparaître que plus de 98% de l'inertie est expliquée.

L'arbre de classification obtenu à l'issue de l'algorithme est présenté à la figure 4.3 :

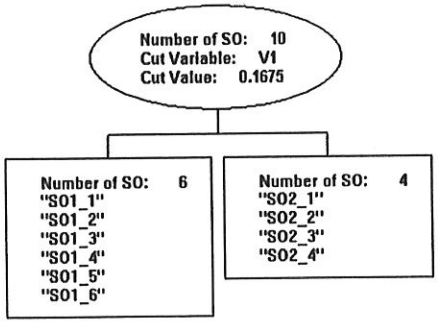

FIG. 4.3 – *Arbre de classification.*

Remarque : La sortie de la méthode DIV est extrêmement simple à interpréter. Elle est claire, succincte et complète.

4.1.4 La méthode SCLASS

Les résultats de la méthode SCLASS sont repris ci-dessous.

```
=========================
Split of the node :    1
=========================

Number of Symbolic objects in the node :   10pt
---------------------------------------

Criteria of cut :
-----------------

Cut variable :(  1) V1
Cut value :   0.17
Smoothing parameter CENTER : 5.66
Smoothing parameter LENGTH : 0.24

Rule : if value of i < 0.17 -> the SO i is in the left node
       if value of i > 0.17 -> the SO i is in the right node
```

La coupure du noeud 1 se fait suivant la variable V1. Elle permet de séparer l'ensemble des objets en deux classes.

À l'issue de cette première coupure, nous obtenons les deux noeuds suivants :

```
Node :   2    Cardinal :   6pt
================================
(0)  SO1_1
(1)  SO1_2
(2)  SO1_3
(3)  SO1_4
(4)  SO1_5
(5)  SO1_6

Node :   3    Cardinal :   4pt
================================
(6)  SO2_1
(7)  SO2_2
(8)  SO2_3
(9)  SO2_4
```

De même que pour les méthodes SCLUST et DIV, la partition des objets en deux classes obtenue à l'aide de la méthode SCLASS correspond bien à la partition naturelle des données.

La division successive des noeuds se poursuit jusqu'à ce que l'effectif minimal fixé par l'utilisateur soit atteint pour chaque noeud. Ici, nous avons fixé l'effectif minimal à 3. L'arbre de classification obtenu à la fin de l'algorithme est représenté à la figure 4.4 :

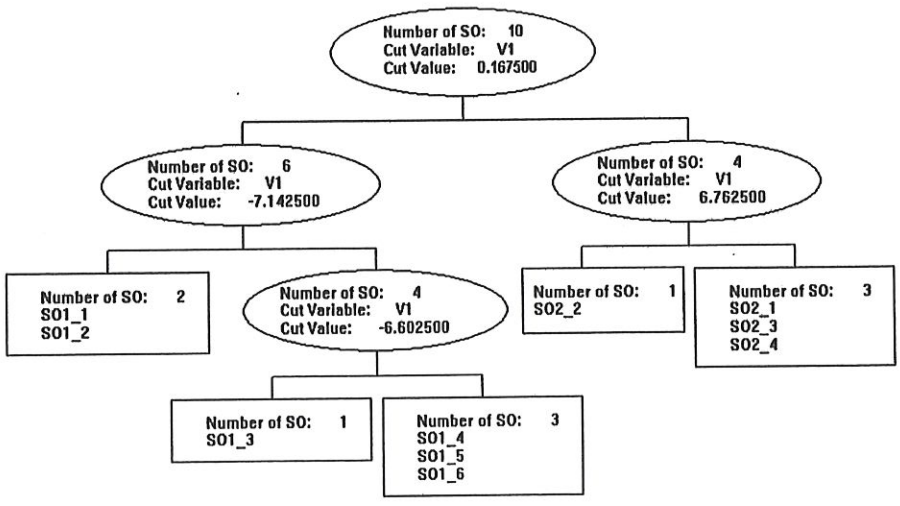

FIG. 4.4 – *Arbre de classification.*

Normalement, une fois le processus de coupure terminé, il est prévu que l'arbre de classification obtenu soit élagué. Malheureusement, dans la sortie des résultats de la méthode SCLASS, aucun indicateur ne signale le bon déroulement de la procédure d'élagage de l'arbre, ni même les branches élaguées. Ceci nous pousse à croire que l'arbre de classification présenté n'a pas été élagué ou que la procédure d'élagage ne fonctionne pas dans SODAS.

4.1.5 La méthode HIPYR

Nous avons testé la méthode HIPYR avec les deux critères d'agrégation disponibles, à savoir le critère du degré de généralité et celui de l'incrément du degré de généralité[4].

Les hiérarchies de partitions obtenues avec chacun des critères sont présentées aux figures 4.5 et 4.6 :

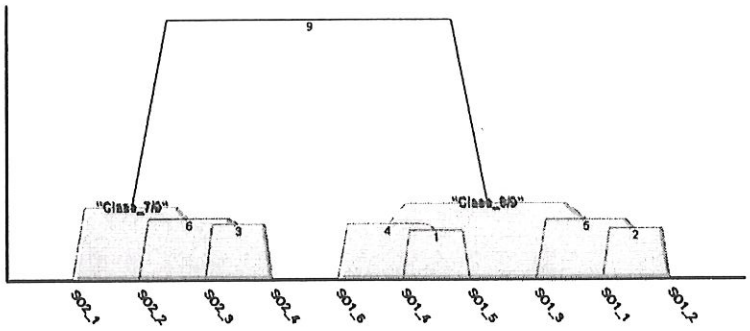

FIG. 4.5 – *Hiérarchie de partitions - Critère du degré de généralité.*

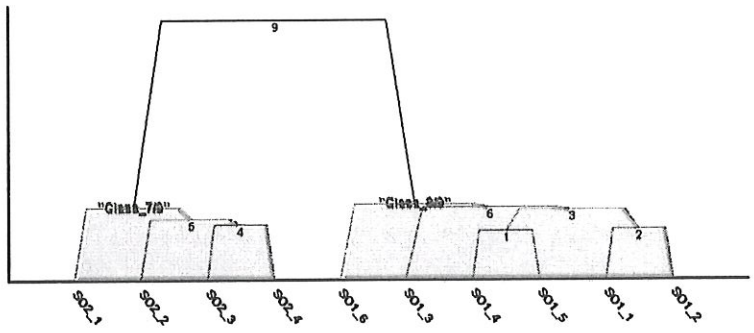

FIG. 4.6 – *Hiérarchie de partitions - Critère de l'incrément du degré de généralité.*

4. Pour pouvoir utiliser les critères du lien simple ou du lien complet, celui de la moyenne ou encore celui du diamètre, il faut travailler à partir de la matrice de dissimilarités entre objets et non pas à partir de l'ensemble des objets symboliques comme nous le faisons.

En examinant les deux dendrogrammes, nous constatons que la partition naturelle des données en deux classes est retrouvée dans les deux cas. Les objets symboliques sont agrégés différemment en fonction du critère d'agrégation utilisé mais la partition finale en deux classes est la même suivant les deux critères.

4.1.6 Conclusion

Les méthodes SCLUST, DIV, SCLASS et HIPYR retrouvent chacune la partition naturelle des données en deux classes.

4.2 Données avec trois classes hypersphériques

4.2.1 Le jeu de données

Le jeu de données que nous étudions dans le cadre cet exemple est composé de 25 objets symboliques décrits par 2 variables intervalles.

Pour créer ce jeu de données, nous avons généré les minima et maxima de chaque intervalle suivant chacune des deux variables de manière à obtenir trois classes hypersphériques bien séparées, à l'aide d'un outil de génération de nombres aléatoires en Excel.

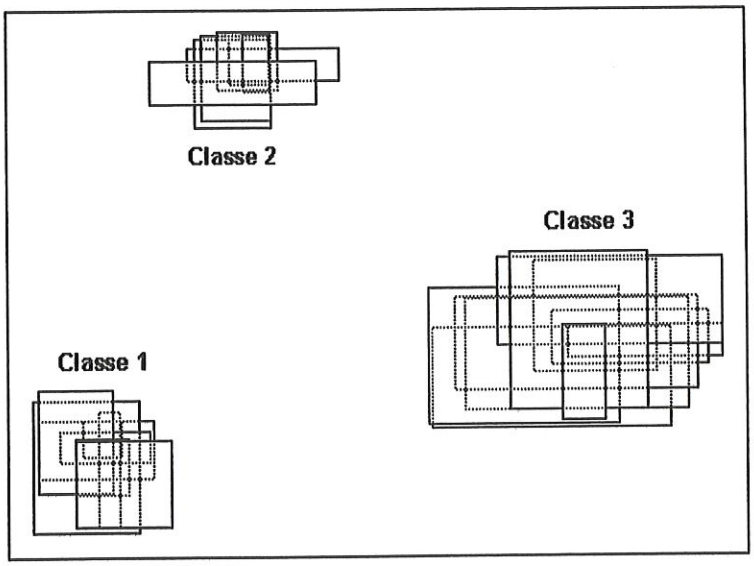

FIG. 4.7 – *Représentation graphique des objets symboliques.*

Étant donnée la répartition les classes, nous nous attendons à ce que les méthodes retrouvent la partition naturelle des objets en trois classes.

Pour clarifier les sorties du logiciel SODAS, nous avons labélisé les objets symboliques comme suit[5] :

- classe 1 : `S01_x` ;
- classe 2 : `S02_y` ;
- classe 3 : `S03_z`.

5. La numérotation des classes indiquée ci-dessous diffère de la numérotation adoptée par SCLUST et DIV.

4.2.2 La méthode SCLUST

Comme pour l'exemple précédent, nous avons testé la méthode avec les trois distances. Pour ce qui est de l'initialisation, nous avons à nouveau opté pour une initialisation par prototypes.

Nous avons réinitialisé 50 fois l'algorithme. La partition optimale des données en trois classes trouvée par SCLUST avec chacune des trois distances est la suivante :

```
Classe :   1  Cardinal :   10
==============================

( 15) S03_1    [1.6]  ( 16) S03_2    [1.2]  ( 17) S03_3    [0.5]
( 18) S03_4    [0.3]  ( 19) S03_5    [1.1]  ( 20) S03_6    [0.5]
( 21) S03_7    [0.5]  ( 22) S03_8    [0.8]  ( 23) S03_9    [1.8]
( 24) S03_10   [1.6]

Classe :   2  Cardinal :    8
==============================

( 0) S01_1    [1.3]  ( 1) S01_2    [1.1]  ( 2) S01_3    [0.6]
( 3) S01_4    [0.8]  ( 4) S01_5    [0.3]  ( 5) S01_6    [1.4]
( 6) S01_7    [1.1]  ( 7) S01_8    [1.4]

Classe :   3  Cardinal :    7
==============================

( 8) S02_1    [2.1]  ( 9) S02_2    [1.8]  ( 10) S02_3    [0.6]
( 11) S02_4   [0.4]  ( 12) S02_5   [0.2]  ( 13) S02_6    [0.7]
( 14) S02_7   [1.1]
```

Comme nous pouvons le constater, il s'agit bien de la partition naturelle des données en trois classes.

Les prototypes des classes[6] pour chacune des deux variables sont représentés à la figure 4.8.

Nous constatons que les prototypes des classes sont bien séparés selon les deux variables. La variable V1 permet de discriminer la classe 1 des classes 2 et 3. La variable V2 permet quant à elle de discriminer la classe 2 des classes 1 et 3.

6. Avec la distance de Hausdorff.

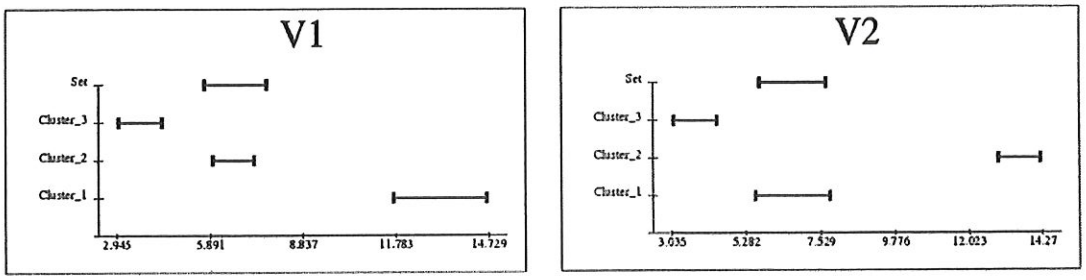

FIG. 4.8 – *Prototypes des deux classes pour les variables* V1 *et* V2.

La description de la partition optimale trouvée par SCLUST est reprise dans le tableau 4.2 :

Distances	Nombre de fois où la partition naturelle en 3 classes est retrouvée	Valeur initiale du critère	Valeur finale du critère	Pourcentage d'inertie expliquée par la partition
Hausdorff	36	215.12	38.10	82.29
$\mathcal{L}^1$	39	410.51	58.19	85.83
$\mathcal{L}^2$	35	2169.01	51.30	97.63

TAB. 4.2 – *Résultats de la méthode* SCLUST.

En regardant le tableau, nous constatons que la partition naturelle des données en trois classes est retrouvée entre 35 et 39 fois sur 50 selon la distance utilisée.

En assimilant les objets aux prototypes des classes les plus proches, nous parvenons à expliquer plus de 80% de l'inertie contenue dans les données. Avec la distance $\mathcal{L}^2$, nous atteignons les 97% d'inertie expliquée.

4.2.3 La méthode DIV

La hiérarchie de partitions produite par la méthode DIV est la suivante :

```
PARTITION IN 2 CLUSTERS :
-------------------------

Cluster 1 (n=15) : S01_1 S01_2 S01_3 S01_4 S01_5 S01_6 S01_7 S01_8
                   S02_1 S02_2 S02_3 S02_4 S02_5 S02_6 S02_7

Cluster 2 (n=10) : S03_1 S03_2 S03_3 S03_4 S03_5 S03_6 S03_7 S03_8
                   S03_9 S03_10

Explicated inertia : 63.868741
```

Nous constatons ici que la classe hypersphérique notée `Classe 3` dans l'espace de représentation (figure 4.7) est séparée des deux autres classes.

Le critère de décision pour la coupure en deux classes se base sur la variable V1. La valeur de coupure associée est de 11.16.

```
Cluster 1 :
   IF    1- [V1 <= 11.160000] IS TRUE

Cluster 2 :
   IF    1- [V1 <= 11.160000] IS FALSE
```

La seconde coupure engendre la partition suivante :

```
PARTITION IN 3 CLUSTERS :
-------------------------

Cluster 1 (n=8)  : SO1_1 SO1_2 SO1_3 SO1_4 SO1_5 SO1_6 SO1_7 SO1_8

Cluster 2 (n=10) : SO3_1 SO3_2 SO3_3 SO3_4 SO3_5 SO3_6 SO3_7 SO3_8
                   SO3_9 SO3_10

Cluster 3 (n=7)  : SO2_1 SO2_2 SO2_3 SO2_4 SO2_5 SO2_6 SO2_7

Explicated inertia : 96.854635
```

Comme nous pouvons le constater, cette partition en trois classes correspondant bien à la partition naturelle des objets symboliques.

La division de l'ensemble des objets symboliques en trois classes permet d'expliquer plus de 96% de l'inertie contenue dans les données.

Les trois classes obtenues peuvent être décrites de la manière suivante :

```
Cluster 1 :
   IF    2- [V2 <= 8.762500] IS TRUE
   AND   1- [V1 <= 11.16000] IS TRUE

Cluster 2 :
   IF    1- [V1 <= 11.16000] IS FALSE

Cluster 3 :
   IF    2- [V2 <= 8.762500] IS FALSE
   AND   1- [V1 <= 11.16000] IS TRUE
```

La seconde coupure se fait suivant la variable V2. Elle permet de séparer les deux classes notées `Classe 1` et `Classe 2` dans l'espace de représentation (figure 4.7).

L'arbre de classification obtenu à la fin de l'algorithme est présenté à la figure 4.9 :

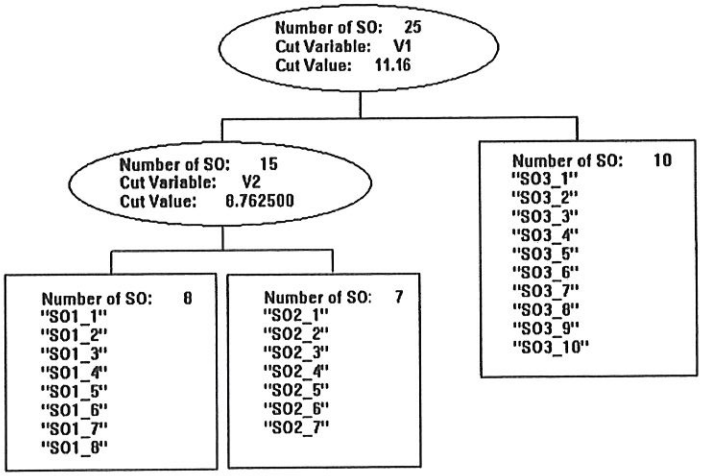

FIG. 4.9 – *Arbre de classification.*

Remarque : Le parcours successif des branches de l'arbre nous fournit une première description des classes. Nous savons, par exemple, que les milieux des intervalles de la classe notée `Classe 1` dans l'espace de représentation (figure 4.7) suivant la variable `V1` sont inférieurs à 11.16, tandis que milieux des intervalles de cette même classe suivant la variable `V2` sont inférieurs à 8.7625.

4.2.4 La méthode SCLASS

Les résultats obtenus par la méthode SCLASS sont les suivants :

```
=========================
Split of the node :    1
=========================

Number of Symbolic objects in the node :   25pt
----------------------------------------

Criteria of cut :
-----------------

Cut variable : V2
Cut value : 6.91

Smoothing parameter CENTER : 2.70
Smoothing parameter LENGTH : 0.25

Rule : if value of i < 6.91 -> the SO i is in the left node
       if value of i > 6.91 -> the SO i is in the right node
```

La première coupure s'effectue suivant la variable V2. Elle engendre les deux noeuds suivants :

```
Node :   2    Cardinal : 15pt
================================
  (0)  S01_1
  (1)  S01_2
  (2)  S01_3
  (3)  S01_4
  (4)  S01_5
  (5)  S01_6
  (6)  S01_7
  (7)  S01_8
 (15)  S03_1
 (16)  S03_2
 (17)  S03_3
 (18)  S03_4
 (22)  S03_8
 (23)  S03_9
 (24)  S03_10

Node :   3    Cardinal : 10pt
================================
  (8)  S02_1
  (9)  S02_2
 (10)  S02_3
 (11)  S02_4
 (12)  S02_5
 (13)  S02_6
 (14)  S02_7
 (19)  S03_5
 (20)  S03_6
 (21)  S03_7
```

En analysant la composition de ces deux noeuds, nous constatons que la première coupure engendre d'ores et déjà une séparation des objets symboliques au sein de la classe notée `Classe 3` dans l'espace de représentation (figure 4.7).

```
=========================
Split of the node :    2
=========================

Number of Symbolic objects in the node:     15pt
---------------------------------------

Criteria of cut :
-----------------

Cut variable :(  2) V2
Cut value : 5.14

Smoothing parameter CENTER :   1.05
Smoothing parameter LENGTH :   0.25

Rule : if value of i <  5.14 -> the SO i is in the left node
       if value of i >  5.14 -> the SO i is in the right node
```

À la deuxième étape de l'algorithme, c'est au tour du noeud 2 à être coupé. La coupure du noeud se fait à nouveau suivant la variable V2.

Cette deuxième coupure donne lieu aux deux noeuds suivants :

```
Node :   4    Cardinal :  8pt
================================
 (0)  SO1_1
 (1)  SO1_2
 (2)  SO1_3
 (3)  SO1_4
 (4)  SO1_5
 (5)  SO1_6
 (6)  SO1_7
 (7)  SO1_8

Node :   5    Cardinal :  7pt
================================
 (15)  SO3_1
 (16)  SO3_2
 (17)  SO3_3
 (18)  SO3_4
 (22)  SO3_8
 (23)  SO3_9
 (24)  SO3_10
```

Comme nous pouvons le constater, la coupure du noeud 2 permet de scinder une partie des objets symboliques de la classe notée `Classe 3` de ceux de la classe notée `Classe 1` dans l'espace de représentation (figure 4.7).

Les divisions des noeuds se succèdent jusqu'à ce que l'effectif minimal fixé par l'utilisateur soit atteint pour chaque noeud. Pour cet exemple, l'effectif minimal d'un noeud est fixé arbitrairement à 6. Une fois l'algorithme terminé, nous obtenons l'arbre de classification présenté à la figure 4.10.

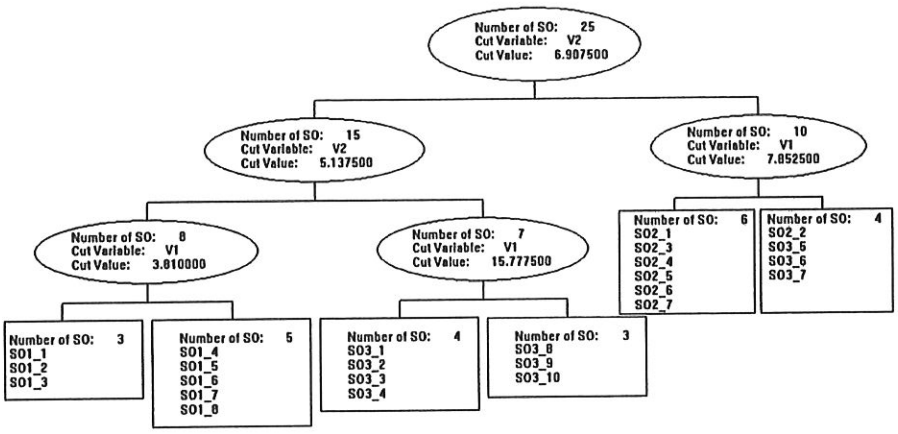

FIG. 4.10 – *Arbre de classification.*

En parcourant attentivement l'arbre de classification, nous observons que les coupures successives des noeuds permettent d'isoler les objets symboliques appartenant à une même classe. À la fin de l'algorithme, la majorité des objets d'une même classe sont regroupés ; seul l'objet `SO2_2` se trouve encore en présence des objets `SO3_5`, `SO3_6` et `SO3_7`. La classe notée `Classe 1` dans l'espace de représentation (figure 4.7) est retrouvée par l'algorithme.

Même si la partition obtenue par la méthode ne correspond pas à la partition naturelle des données en trois classes, nous pouvons dire que les résultats ne sont pas totalement mauvais.

4.2.5 La méthode HIPYR

La hiérarchie de partitions produite par la méthode HIPYR avec le critère du degré de généralité est représentée à la figure 4.11 :

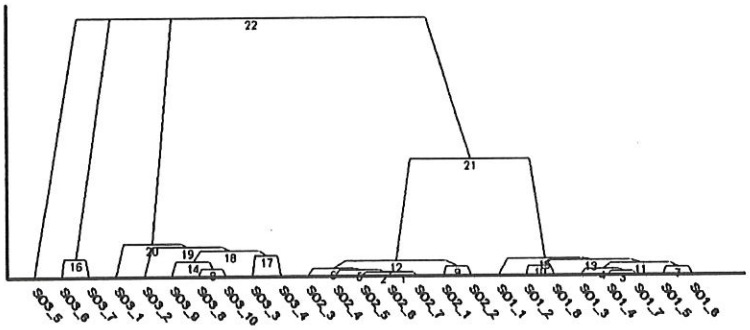

FIG. 4.11 – *Hiérarchie de partitions - Critère du degré de généralité.*

Ce dendrogramme peut nous sembler quelque peu surprenant dans la mesure où quatre classes sont réunies à la dernière étape. Ceci nous empêche d'ailleurs de retrouver la partition naturelle des données en trois classes.

Au cours de cette dernière étape, l'algorithme ne parvient pas à trouver deux classes qui satisfont aux conditions d'agrégation. L'algorithme tente donc d'en trouver plusieurs à la fois comme c'est le cas ici.

Les résultats obtenus par la méthode sont assez positifs. En parcourant le dendrogramme, nous pouvons constater que la structure interne des données est bien retrouvée. L'algorithme ne regroupe pas des objets provenant de classes différentes et parvient à retrouver les deux classes notées `Classe 1` et `Classe 2` dans l'espace de représentation (figure 4.7).

La hiérarchie de partitions produite par la méthode avec le critère de l'incrément du degré de généralité est représentée par le dendrogramme de la figure 4.12.

Cette fois, nous pouvons constater que la partition naturelle des objets en trois classes est bien retrouvée par la méthode.

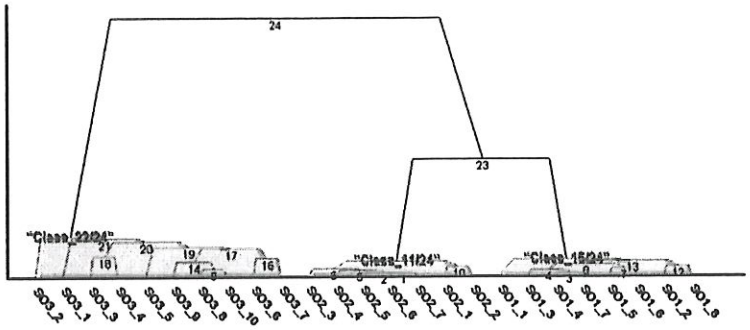

FIG. 4.12 – *Hiérarchie de partitions - Critère d'incrément du degré de généralité.*

4.2.6 Conclusion

Dans cet exemple, la partition naturelle des données en trois classes n'est pas retrouvée par toutes les méthodes.

Les méthodes SCLUST, DIV et HIPYR utilisée avec le critère de l'incrément du degré de généralité permettent de retrouver la partition naturelle des données en trois classes. HIPYR utilisée avec le critère du degré de généralité ne permet pas de retrouver cette partition en trois classes, de même que SCLASS qui ne retrouve qu'une partie de la structure naturelle des données.

4.3 Données avec deux classes allongées

4.3.1 Le jeu de données

Le jeu de données que nous étudions dans le cadre cet exemple est composé de 20 objets symboliques décrits par 2 variables intervalles.

Pour créer ce jeu de données, nous avons généré les minima et maxima de chaque intervalle suivant chacune des deux variables de manière à obtenir deux classes allongées, à l'aide d'un outil de génération de nombres aléatoires en Excel.

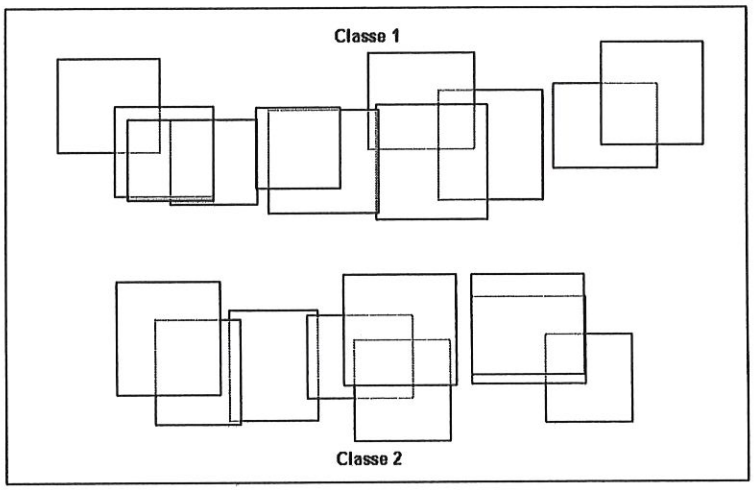

FIG. 4.13 – *Représentation graphique des objets symboliques.*

Pour clarifier les sorties du logiciel SODAS, nous avons labélisé les objets symboliques comme suit :

- classe 1 : `S01_x` ;
- classe 2 : `S02_y`.

4.3.2 La méthode SCLUST

Pour ce qui est de l'initialisation, nous avons opté pour une initialisation par prototypes.

Comme pour les exemples précédents, nous avons réinitialisé 50 fois l'algorithme. Les résultats obtenus sont décrits ci-dessous.

La partition optimale trouvée avec la distance de Hausdorff et la distance $\mathcal{L}^1$ est la suivante :

```
Classe :   1  Cardinal :     11
================================

( 0) S01_1     [1.6]  ( 1) S01_2     [1.2]  ( 2) S01_3     [1.2]
( 3) S01_4     [0.9]  ( 4) S01_5     [0.4]  ( 5) S01_6     [0.3]
( 6) S01_7     [0.6]  ( 7) S01_8     [0.6]  ( 8) S01_9     [0.8]
( 9) S01_10    [1.5]  ( 10) S01_11   [2.0]

Classe :   2  Cardinal :      9
================================

( 11) S02_1    [1.7]  ( 12) S02_2    [1.5]  ( 13) S02_3    [0.9]
( 14) S02_4    [0.2]  ( 15) S02_5    [0.3]  ( 16) S02_6    [0.4]
( 17) S02_7    [1.3]  ( 18) S02_8    [1.2]  ( 19) S02_9    [1.6]
```

Comme nous pouvons le constater, il s'agit bien de la partition naturelle des objets en deux classes.

Les prototypes des classes[7] pour chacune des deux variables sont représentés à la figure 4.14.

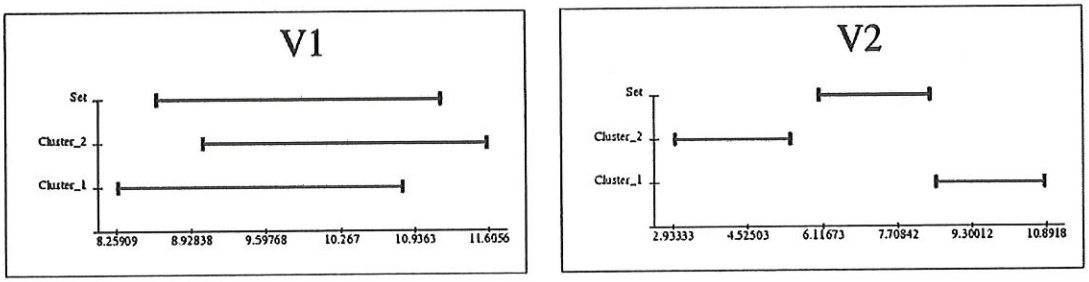

FIG. 4.14 – *Prototypes des deux classes pour les variables* V1 *et* V2.

Les prototypes des classes sont bien séparés suivant la variable V2. Ceci souligne le pouvoir discriminant de cette variable.

7. Avec la distance $\mathcal{L}^1$.

La description de la partition optimale trouvée par SCLUST pour les distances de Hausdorff et $\mathcal{L}^1$ est reprise dans le tableau 4.3 :

Distances	Nombre de fois où la partition naturelle en 2 classes est retrouvée	Valeur initiale du critère	Valeur finale du critère	Pourcentage d'inertie expliquée par la partition
Hausdorff	22	126.86	84.11	33.70
$\mathcal{L}^1$	18	251.90	163.52	35.09

TAB. 4.3 – *Résultats de la méthode* SCLUST.

Au vu de ce tableau, nous remarquons que la partition naturelle des données en deux classes n'est pas retrouvée très souvent, que ce soit avec la distance de Hausdorff ou la distance $\mathcal{L}^1$.

En assimilant les objets aux prototypes des classes les plus proches, le pourcentage d'inertie expliquée n'atteint que 33% pour la distance de Hausdorff et 35% pour la distance $\mathcal{L}^1$. Ce qui n'est pas beaucoup en soi.

Discussion supplémentaire

Nous avons constaté que la partition naturelle des données en deux classes allongées est également retrouvée par SCLUST avec la distance $\mathcal{L}^2$ si les classes allongées sont quelque peu écartées. En effet, la partition naturelle des données en deux classes n'était pas retrouvée avec la distance euclidienne sur les données originales.

Nous avons éloigné les deux classes allongées de 2 unités. La description de la partition optimale trouvée par SCLUST avec chacune des distances est reprise dans le tableau 4.4 :

Distances	Nombre de fois où la partition naturelle en 2 classes est retrouvée	Valeur initiale du critère	Valeur finale du critère	Pourcentage d'inertie expliquée par la partition
Hausdorff	28	144.86	84.11	41.94
$\mathcal{L}^1$	24	291.50	163.52	43.90
$\mathcal{L}^2$	18	1264.20	702.39	44.44

TAB. 4.4 – *Résultats de la méthode* SCLUST.

Même si la partition optimale trouvée par SCLUST correspond bien à la partition naturelle des données en deux classes, nous constatons que la fréquence avec laquelle la partition naturelle est retrouvée est relativement faible. Ceci justifie la nécessité de procéder à plusieurs initialisations de l'algorithme.

4.3.3 La méthode Div

La partition de l'ensemble des objets en deux classes produite par la méthode Div est la suivante :

```
PARTITION IN 2 CLUSTERS :
------------------------:

Cluster 1 (n=9)  : S01_1 S01_2 S01_3 S01_4 S01_5 S01_6 S02_1 S02_2
                   S02_3

Cluster 2 (n=11) : S01_7 S01_8 S01_9 S01_10 S01_11 S02_4 S02_5
                   S02_6 S02_7 S02_8 S02_9

Explicated inertia : 50.607735
```

Comme nous pouvons le constater, cette partition ne correspond pas à la partition naturelle des données en deux classes.

La description de la partition est donnée ci-dessous. Elle illustre la manière dont les objets sont classés.

```
Cluster 1 :
   IF    1- [V1 <= 9.265000] IS TRUE

Cluster 2 :
   IF    1- [V1 <= 9.265000] IS FALSE
```

La coupure de l'ensemble des objets se fait suivant la variable V1, ce qui empêche de retrouver les deux classes allongées.

L'arbre de classification obtenu à l'issue de l'algorithme est présenté à la figure 4.15 :

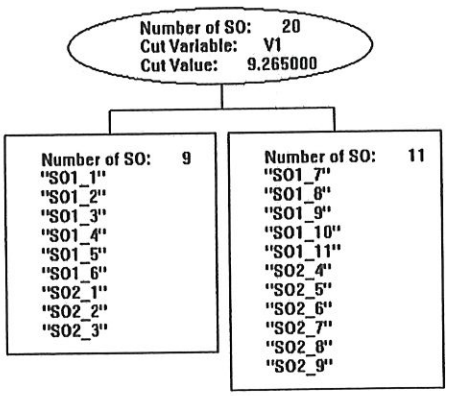

FIG. 4.15 – *Arbre de classification.*

Si la partition optimale n'est pas retrouvée ici, c'est parce que la méthode est basée sur un critère d'inertie. La coupure de l'ensemble des objets se fait suivant la variable V1 car la perte d'inertie est maximale suivant cette variable.

Discussion supplémentaire

En écartant quelque peu les classes allongées, nous avons constaté que la partition naturelle des données était retrouvée par l'algorithme.

Nous avons éloigné les classes de 2 unités. La description de la partition trouvée est la suivante :

```
Cluster 1 :
   IF    1- [V2 <= 8.022500] IS TRUE

Cluster 2 :
   IF    1- [V2 <= 8.022500] IS FALSE
```

Comme nous pouvons le constater, la coupure se fait à présent suivant la variable V2. Elle permet cette fois de séparer correctement les classes allongées.

L'arbre de classification obtenu à la fin de l'algorithme est présenté à la figure 4.16 :

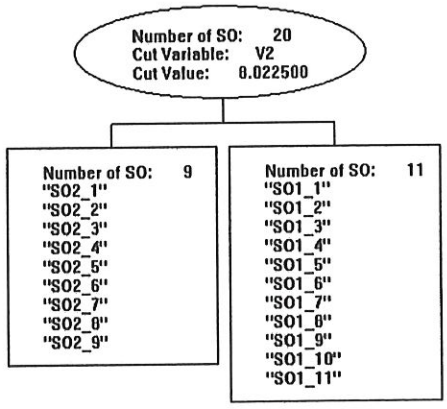

FIG. 4.16 – *Arbre de classification.*

4.3.4 La méthode SCLASS

La construction de l'arbre de classification se déroule de la manière suivante :

```
=========================
Split of the node :    1
=========================

Number of Symbolic objects in the node:  20pt
---------------------------------------

Criteria of cut :
-----------------

Cut variable :(  1) V1
Cut value :  14.83

Smoothing parameter CENTER :   2.14
Smoothing parameter LENGTH :   0.08

Rule : if value of i <   14.83 -> the SO i is in the left node
       if value of i >   14.83 -> the SO i is in the right node
```

Le premier noeud est coupé suivant la variable V1. Les noeuds résultant de cette coupure sont décrits ci-dessous :

```
Node :   2    Cardinal :  17pt
================================
  (0)  SO1_1
  (1)  SO1_2
  (2)  SO1_3
  (3)  SO1_4
  (4)  SO1_5
  (5)  SO1_6
  (6)  SO1_7
  (7)  SO1_8
  (8)  SO1_9
 (11)  SO2_1
 (12)  SO2_2
 (13)  SO2_3
 (14)  SO2_4
 (15)  SO2_5
 (16)  SO2_6
 (17)  SO2_7
 (18)  SO2_8

Node :   3    Cardinal :   3pt
================================
  (9)  SO1_10
 (10)  SO1_11
 (19)  SO2_9
```

À ce stade de l'algorithme, nous constatons que la partition en deux classes trouvée par SCLASS ne correspond pas à la partition naturelle des données.

Cependant, si nous poursuivons l'analyse de la construction de la hiérarchie, nous remarquons que la seconde coupure divise le noeud 2 suivant la variable V2 et que la partition naturelle des données est donc en partie retrouvée.

La coupure du noeud 2 engendre les deux noeuds suivants :

```
Node :   4    Cardinal :   8pt
================================
 (11)  S02_1
 (12)  S02_2
 (13)  S02_3
 (14)  S02_4
 (15)  S02_5
 (16)  S02_6
 (17)  S02_7
 (18)  S02_8

Node :   5    Cardinal :   9pt
================================
 (0)  S01_1
 (1)  S01_2
 (2)  S01_3
 (3)  S01_4
 (4)  S01_5
 (5)  S01_6
 (6)  S01_7
 (7)  S01_8
 (8)  S01_9
```

Comme nous pouvons le constater, la coupure du noeud 2 permet de séparer les objets symboliques en deux classes de sorte que :

- dans le noeud 4, nous retrouvons uniquement des objets qui appartiennent à la classe notée `Classe 2` dans l'espace de représentation (figure 4.13);
- dans le noeud 5, nous retrouvons uniquement des objets qui appartiennent à la classe notée `Classe 1` dans l'espace de représentation (figure 4.13).

La partition des données en trois classes permet donc de retrouver une partie de la structure allongées des données.

L'arbre de classification obtenu à la fin de l'algorithme est présenté à la figure 4.17 :

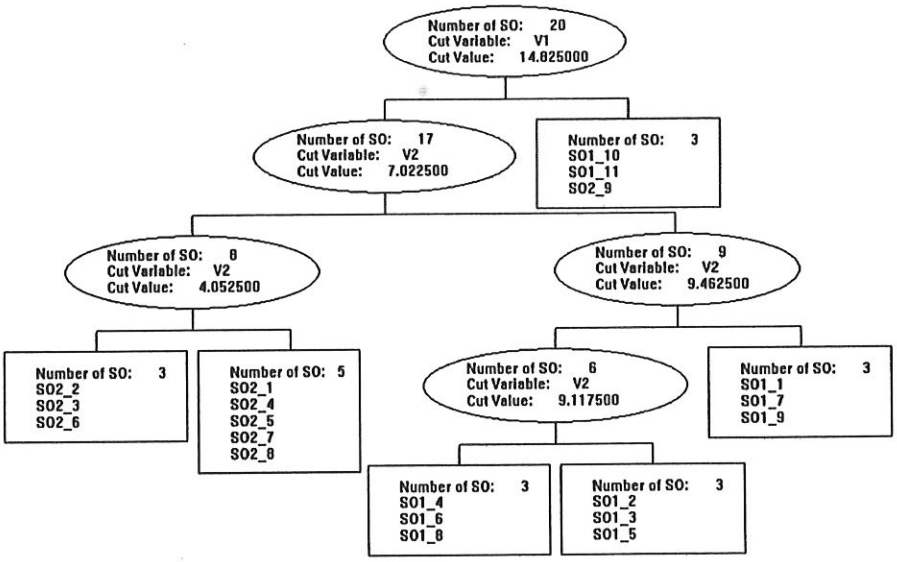

FIG. 4.17 – *Arbre de classification.*

Remarque : Il est surprenant de voir que l'algorithme de SCLASS ne parvient pas à retrouver la partition naturelle des données, et ce même lorsque les classes allongées sont légèrement écartées l'une de l'autre. Pourtant, le principe de classification des données sous-jacent à cette méthode devrait permettre de retrouver les deux classes allongées.

4.3.5 La méthode HIPYR

En appliquant la méthode HIPYR avec les critères du degré de généralité et de l'incrément du degré de généralité, nous obtenons les hiérarchies de partitions présentées aux figures 4.18 et 4.19.

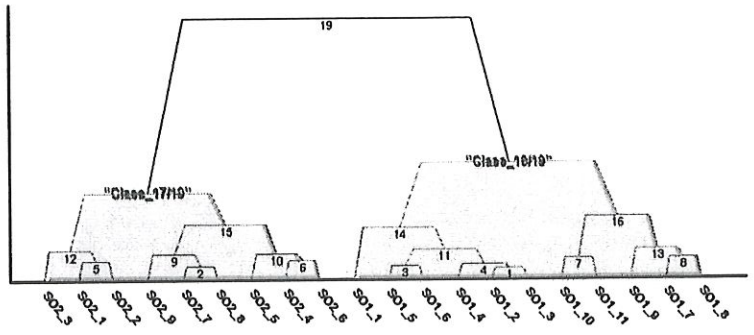

FIG. 4.18 – *Hiérarchie de partitions - Critère du degré de généralité.*

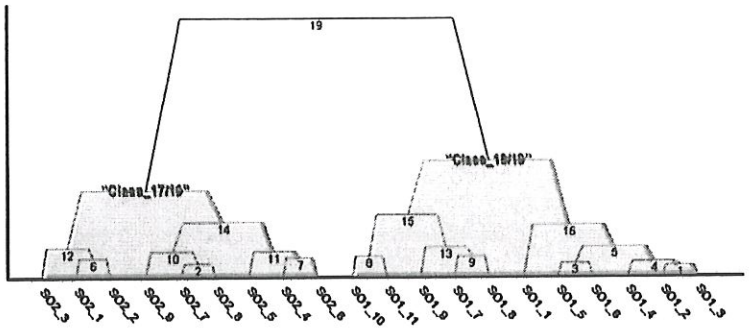

FIG. 4.19 – *Hiérarchie de partitions - Critère de l'incrément du degré de généralité.*

Nous constatons sur ces deux dendrogrammes que la méthode HIPYR permet de retrouver la partition naturelle des données. Dans les deux cas, nous retrouvons bien les deux classes allongées.

4.3.6 Conclusion

Après avoir analysé les résultats produits par chacune des méthodes, nous en sommes arrivés à la conclusion que les méthodes SCLUST, DIV et HIPYR retrouvent une structure allongée dans les données. Toutefois, les méthodes SCLUST avec utilisation de la distance euclidienne et DIV ne permettent de retrouver les deux classes allongées que lorsque celles-ci sont suffisamment écartées.

Pour ce qui est de la méthode SCLASS, nous sommes assez surpris de voir que l'algorithme ne parvient pas à retrouver les classes allongées. La partition en trois classes trouvée par l'algorithme se rapproche toutefois fortement de la partition naturelle des données.

4.4 Données avec deux classes emboîtées

4.4.1 Le jeu de données

Le jeu de données que nous traitons dans le cadre ce dernier exemple est composé de 16 objets symboliques décrits par une seule variable intervalle.

Les objets symboliques ont été générés aléatoirement de manière à obtenir deux classes emboîtées. Les centres des intervalles sont compris entre 0 et 5, tandis que leur longueur est comprise entre 1 et 2 pour une classe et entre 11 et 12 pour l'autre.

Pour clarifier les sorties du logiciel SODAS, nous avons labélisé les objets symboliques comme suit :

- classe 1 : `S01_x` ;
- classe 2 : `S02_y`.

4.4.2 La méthode SCLUST

La partition optimale en deux classes trouvée par la méthode SCLUST avec chacune des trois distances est la suivante :

```
Classe :   1  Cardinal :      7
================================

( 0) S01_1    [0.2]  ( 1) S01_2    [1.5]  ( 2) S01_3    [0.0]
( 3) S01_4    [0.9]  ( 4) S01_5    [0.9]  ( 5) S01_6    [1.3]
( 6) S01_7    [2.2]

Classe :   2  Cardinal :      9
================================

( 7) S02_1    [0.5]  ( 8) S02_2    [6.2]  ( 9) S02_3    [0.2]
( 10) S02_4   [0.0]  ( 11) S02_5   [0.6]  ( 12) S02_6   [0.7]
( 13) S02_7   [0.1]  ( 14) S02_8   [0.0]  ( 15) S02_9   [0.8]
```

Les prototypes des classes pour la variable `V1` sont présentés à la figure 4.20.

Au vu de ce graphique, nous constatons que la longueur du prototype de la classe 2 est nettement supérieure à celui de la classe 1. Cela vient du fait que les intervalles de la classe 1 sont emboîtés dans les intervalles de la classe 2.

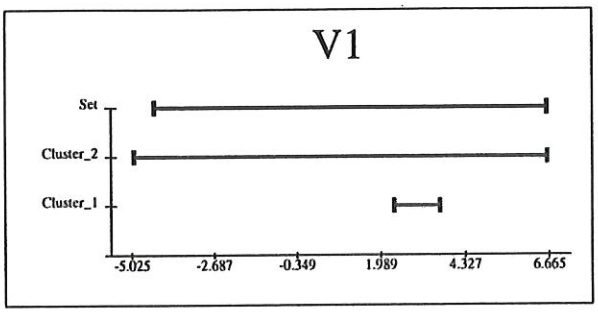

FIG. 4.20 – *Prototypes des classes.*

La description de la partition optimale trouvée par SCLUST est reprise dans le tableau 4.5 :

Distances	Nombre de fois où la partition naturelle en 2 classes est retrouvée	Valeur initiale du critère	Valeur finale du critère	Pourcentage d'inertie expliquée par la partition
Hausdorff	50	58.26	20.60	64.63
$\mathcal{L}^1$	48	79.81	39.54	50.46
$\mathcal{L}^2$	50	294.35	74.70	74.72

TAB. 4.5 – *Résultats de la méthode* SCLUST.

Nous constatons au vu de ce tableau que la structure emboîtée des données est pratiquement retrouvée à chaque fois par SCLUST avec chacune des trois distances.

En assimilant les objets aux prototypes des classes les plus proches, une part considérable de l'inertie est expliquée.

Si la méthode SCLUST parvient à retrouver la partition des petits et des grands intervalles, c'est dû au fait que :

- les centres des intervalles ont été générés sur un petit segment (entre 0 et 5) ;
- il y a une différence de longueur assez importante entre les petits et les grands intervalles.

Pour que les classes emboîtées puissent être retrouvées, il faut que la dissimilarité entre des intervalles de même type (petits ou grands intervalles) soit inférieure à celle entre des intervalles de type différents.

4.4.3 La méthode Div

La partition de l'ensemble des objets en deux classes produite par la méthode Div est la suivante :

```
PARTITION IN 2 CLUSTERS :
-------------------------

Cluster 1 (n=11) : SO1_1 SO1_2 SO1_7 SO2_1 SO2_3 SO2_4
                   SO2_5 SO2_6 SO2_7 SO2_8 SO2_9

Cluster 2 (n=5)  : SO1_3 SO1_4 SO1_5 SO1_6 SO2_2

Explicated inertia : 55.248414
```

Comme nous pouvons le constater, cette partition ne correspond pas à la partition naturelle des données en deux classes.

La description de la partition ci-dessous permet d'illustrer la manière dont les objets ont été séparés.

```
Cluster 1 :
   IF    1- [V1 <= 2.510000] IS TRUE

Cluster 2 :
   IF    1- [V1 <= 2.510000] IS FALSE
```

La coupure suivant l'unique variable V1 ne permet pas de retrouver la classe des petits et des grands intervalles. Ceci est simplement dû au fait que les centres des intervalles sont répartis aléatoirement entre 0 et 5.

L'arbre de classification obtenu est présenté à la figure 4.21 :

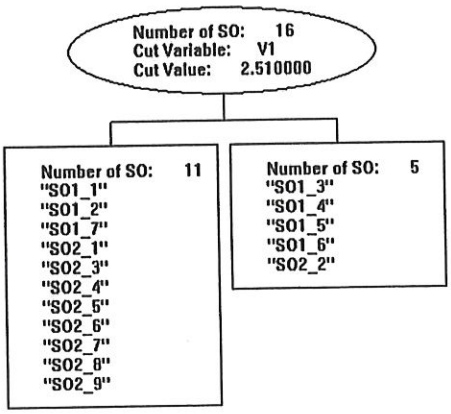

FIG. 4.21 – *Arbre de classification.*

4.4.4 La méthode SCLASS

La hiérarchie de partitions produite par l'algorithme est décrite ci-dessous :

```
Split of the node :    1
=========================

Number of Symbolic objects in the node:     16pt
--------------------------------------

Criteria of cut :
-----------------

Cut variable :(  1) V1
Cut value :   3.65
Smoothing parameter CENTER : 1.12
Smoothing parameter LENGTH : 2.17

Rule : if value of i < 3.65 -> the SO i is in the left node
       if value of i > 3.65 -> the SO i is in the right node
```

À l'issue de la coupure du noeud 1, nous obtenons les deux noeuds suivants :

```
Node :   2    Cardinal :    12pt
====================================
(0)  SO1_1
(1)  SO1_2
(2)  SO1_3
(6)  SO1_7
(7)  SO2_1
(9)  SO2_3
(10) SO2_4
(11) SO2_5
(12) SO2_6
(13) SO2_7
(14) SO2_8
(15) SO2_9

Node :   3    Cardinal :     4pt
====================================
(3)  SO1_4
(4)  SO1_5
(5)  SO1_6
(8)  SO2_2
```

La partition en deux classes produite par SCLASS ne correspond pas à la partition naturelle des données. Étant donné que les intervalles des deux classes se distinguent par leur longueur et non par leur centre, le critère de coupure de SCLASS ne permet pas la séparation de classes emboîtées.

L'arbre de classification obtenu une fois le processus de division terminé est présenté à la figure 4.22 :

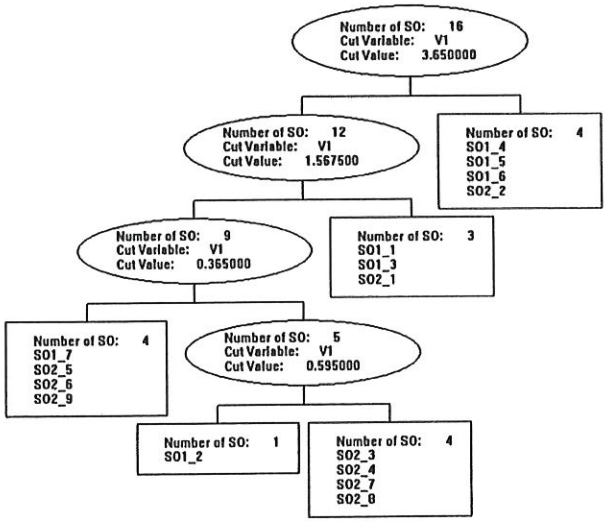

FIG. 4.22 – *Arbre de classification.*

4.4.5 La méthode HIPYR

Les hiérarchies de partitions produites par la méthode HIPYR sont représentées aux figures 4.23 et 4.24.

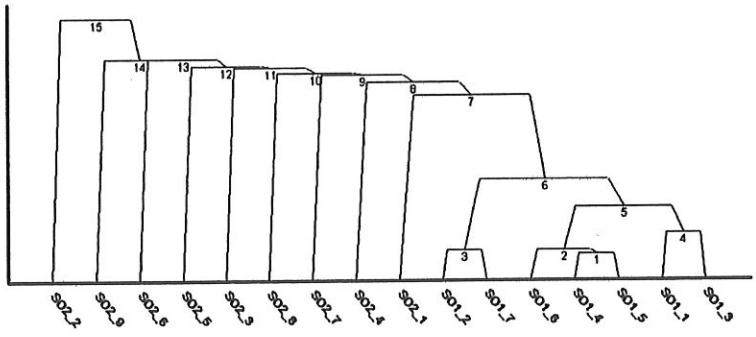

FIG. 4.23 – *Hiérarchie de partitions - Critère du degré de généralité.*

La partition naturelle des objets en deux classes n'est pas retrouvée pour cet exemple. Au cours des premières étapes de l'algorithme, les petits intervalles sont successivement réunis. La classe des petits intervalles est donc formée assez rapidement. Ensuite, à chaque étape de l'algorithme, la classe des intervalles formée jusqu'ici est regroupée avec un singleton.

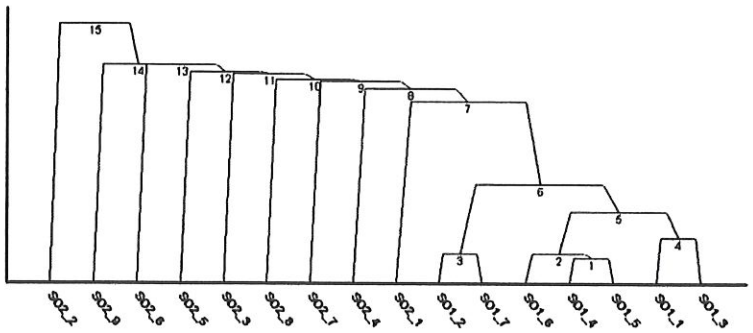

FIG. 4.24 – *Hiérarchie de partitions - Critère de l'incrément du degré de généralité.*

4.4.6 Conclusion

Seule la méthode SCLUST a permis de retrouver les classes des petits et des grands intervalles. Les résultats liés à la partition optimale sont très positifs.

Les partitions en deux classes obtenues par les méthodes DIV et SCLASS ne correspondent pas à la partition naturelle des données. La nature du critère sur lequel ces méthodes sont basées ne permet pas la détection de classes emboîtées.

Les hiérarchies de partitions produites par la méthode HIPYR présentent une structure assez étonnante. Elles ne correspondent pas à la partition naturelle des données. L'algorithme ne parvient pas à identifier la structure emboîtée des données, mais agrège toutefois tous les petits intervalles au cours des premières étapes de l'algorithme.

4.5 Bilan

Dans cette dernière section, nous dressons le bilan de notre étude. Après avoir analysé, les résultats obtenus par les méthodes SCLUST, DIV, SCLASS et HIPYR, voici ce que nous pouvons dire :

- **Données avec deux classes hypersphériques**
  Toutes les méthodes de classification symboliques parviennent à retrouver la partition naturelle des données en deux classes.
- **Données avec trois classes hypersphériques**
  La partition naturelle des données en trois classes est retrouvée par chacune des méthodes, sauf par SCLASS.
- **Données avec deux classes allongées**
  Appliquées à un ensemble de données allongées, les méthodes SCLUST utilisée avec les distances de Haussdorff ou $\mathcal{L}^1$ et HIPYR retrouvent la structure naturelle des données, ainsi que SCLUST utilisée avec la distance euclidienne et DIV lorsque les deux classes allongées sont plus écartées.
- **Données avec deux classes emboîtées**
  Les méthodes de classification symboliques DIV, SCLASS et HIPYR ne parviennent pas à séparer des données emboîtées. La méthode SCLUST retrouve quant à elle la structure naturelle des données.

Remarque : Pour ce qui est de SCLUST, il est préférable de tester la méthode avec chacune des trois distances. En effet, nous avons constaté que les partitions optimales pouvaient différer suivant la distance utilisée.

Chapitre 5

Applications

Dans ce dernier chapitre, deux applications sont exposées. La première traite d'un problème rencontré en logopédie. La seconde concerne les pays membres de l'Union Européenne.

5.1 Application 1 : Le *Voice Handicap Index*

Introduction

En cas de troubles vocaux, une fois le diagnostic de l'oto-rhino-laryngiste posé, le patient subit un examen vocal. Au cours de cet examen, le logopède recueille des données objectives et des données subjectives. Ces deux types de mesures ciblent des aspects différents du trouble vocal. Les mesures objectives ont pour but de quantifier les caractéristiques physiques de la voix (la fréquence, l'intensité, le degré de dysphonie[1]) alors que les mesures subjectives visent plutôt à évaluer la manière dont le patient perçoit son trouble vocal.

Dans la pratique logopédique, le fait de connaître le point de vue du patient, la façon dont il perçoit son trouble vocal et ses conséquences aide le thérapeute à cibler les objectifs de la rééducation.

À l'heure actuelle, il n'existe malheureusement pas encore d'instrument pour quantifier les conséquences psychosociales de troubles vocaux. C'est pourquoi de nombreux chercheurs se sont penchés sur le développement d'outils permettant de mesurer l'impact de ces troubles sur la vie quotidienne.

Dans le cadre de cette étude, nous nous intéressons à un outil en particulier : le *voice handicap index.*

5.1.1 *Voice Handicap Index* (VHI)

Testé à plusieurs reprises, le *voice handicap index* [Jacobson, 1997] s'est révélé être un outil statistique robuste, complet et adaptable à de nombreuses pathologies. Il se présente sous la forme d'un questionnaire qui permet au patient d'évaluer les troubles de sa voix.

Le questionnaire[2] est composé de 30 questions réparties en trois catégories, chacune visant à mesurer un aspect différent du trouble vocal :

- les questions relatives à l'aspect fonctionnel du trouble ont pour objectif de décrire l'impact du trouble vocal sur les activités quotidiennes du sujet.
  Exemple : Mes difficultés de voix limitent ma vie personnelle et sociale.
- les questions relatives à l'aspect émotionnel du trouble visent à mesurer la sensibilité du sujet face à son trouble vocal.
  Exemple : Je suis tendu(e) quand je parle avec d'autres à cause de ma voix.

1. La dysphonie est la transformation anormale du timbre vocal.
2. Le questionnaire complet se trouve en annexe 1.

- les questions relatives à l'aspect physique du trouble cible la perception par le sujet lui-même d'un inconfort au niveau laryngé et les caractéristiques de ses productions vocales.
  Exemple : Je fais beaucoup d'effort pour parler.

Pour chaque question, le sujet est invité à coter, de 0 (jamais) à 4 (toujours), le grade général de la sévérité du trouble. Le résultat global du test d'un sujet se situe donc entre 0 et 120, degré maximal d'impact que le sujet puisse ressentir. Plus le score du sujet est élevé, plus il est affecté par son trouble vocal.

5.1.2 *Voice Handicap Index* adapté à la voix chantée

Comme nous venons de le voir, le *voice handicap index* a été développé pour évaluer la perception qu'a un patient de son trouble vocal.

Les chanteurs constituent un groupe particulier d'individus susceptibles de présenter un trouble vocal dans la mesure où certains perçoivent un trouble en voix parlée uniquement, d'autres principalement en voix chantée.

Des analyses statistiques [Rosen & Mury, 2000] révèlent que les résultats obtenus par les chanteurs au VHI sont significativement inférieurs à ceux obtenus par les non chanteurs. Les chercheurs expliquent cela par le fait que les problèmes de voix rencontrés chez les chanteurs sont assez différents de ceux rencontrés chez les non chanteurs. Ceci dit, même si les résultats des chanteurs au VHI sont faibles, il ne faut pas les ignorer, étant donné l'impact que peut avoir un trouble vocal sur la vie d'un chanteur. Pour mieux répondre aux besoins spécifiques des chanteurs, les chercheurs ont adapté le VHI de Jacobson à la voix chantée.

Le questionnaire du VHI adapté à la voix chantée se trouve ci-dessous. Il consiste en 31 questions. Cette fois aussi, les questions sont réparties en trois catégories[3] fonctionnelle, émotionnelle et physique. Les catégories comptent respectivement 10, 10 et 11 questions.

Questionnaire du VHI adapté à la voix chantée

1. (F) J'ai des difficultés à passer d'un registre à l'autre.
2. (E) Mon problème de voix me tracasse.
3. (P) Je suis à court de souffle quand je chante.
4. (F) J'évite de chanter dans le bruit.
5. (E) Mon problème de voix me gâche parfois le moral.
6. (P) J'ai l'impression que je dois forcer pour chanter.
7. (F) Ma voix passe difficilement au-dessus de l'accompagnement musical.

3. La lettre qui précède chaque question indique la catégorie à laquelle celle-ci appartient.

8. (E) La couleur de ma voix me déplaît (timbre, éclat, mordant, grain,...).
9. (P) Ma voix me lâche par intermittence.
10. (F) Je me sens écarté(e) des "projets" à cause de ma voix.
11. (E) Je trouve que les autres ne comprennent pas mon problème de voix chantée.
12. (P) Ma voix parlée est plus mauvaise après avoir chanté.
13. (F) Mes problèmes de voix entraînent des pertes de revenus.
14. (E) Le fait de chanter me tend, me stresse.
15. (P) Je ressens une gêne ou une douleur dans le larynx quand je chante.
16. (F) Ma voix est plus grave qu'avant.
17. (E) Je me sens diminué(e)/amoindri(e) à cause de ma voix.
18. (P) Le son de ma voix varie au cours d'une même prestation chantée ou d'une répétition ou d'un concert.
19. (F) Ma voix est instable (se dégrade en cours d'émission ou au cours d'un chant).
20. (E) Je suis anxieux (anxieuse) à l'idée de devoir chanter.
21. (P) J'ai du souffle sur la voix.
22. (F) J'adapte difficilement ma voix en fonction des prestations vocales (local, distance, environnement, nombre d'auditeurs, sujet à interpréter, atmosphère).
23. (E) Même quand je ne chante pas, je pense à mon problème de voix.
24. (P) La clarté de ma voix est imprévisible.
25. (F) Même après un échauffement, je n'arrive pas à retrouver une "bonne" voix.
26. (E) Depuis que j'ai un problème de voix, il m'arrive de refuser de chanter.
27. (P) J'ai l'habitude de faire beaucoup d'efforts pour chanter.
28. (F) J'ai des difficultés à traduire mes émotions en chantant.
29. (E) Il m'arrive de perdre espoir quand je pense à mon problème de voix.
30. (P) J'essaie de changer ma façon de chanter pour que ma voix résonne différemment.
31. (P) Ma voix semble cassante et sèche.

En cotant chaque question de 0 (jamais) à 4 (toujours), le chanteur peut évaluer le trouble de sa voix de manière assez précise. Le résultat global du test se situe donc entre 0 et 124. Plus le score obtenu est élevé, plus la perception par le chanteur de son trouble vocal est importante.

C'est le *voice handicap index* adapté à la voix chantée qui nous préoccupe dans le cadre de cette étude.

5.1.3 Contexte de l'étude

Dans le cadre de son mémoire, Charlotte Simon, une étudiante en logopédie de l'Université Catholique de Louvain, s'est intéressée aux troubles vocaux rencontrés chez les chanteurs. L'objectif principal de son étude était de montrer que le questionnaire du VHI est bien adapté pour les chanteurs, c'est-à-dire qu'il permet aux chanteurs d'évaluer leur trouble vocal et d'exprimer l'impact que ce dernier a sur leur quotidien.

Au total, Simon a interrogé 152 sujets. Parmi ceux-ci figurent :
- 37 chanteurs pathologiques ;
- 95 chanteurs sains ;
- 20 non chanteurs sains, appelés contrôles.

Dans le cadre de son étude, elle a demandé aux sujets de remplir le questionnaire du VHI adapté à la voix chantée à deux reprises. Le temps écoulé entre les deux évaluations varie de 5 à 101 jours. Entre les deux évaluations, les interrogés n'ont suivi aucun traitement médical.

Les variables mesurées sur chacun des sujets sont les suivantes[4] :

- `Age` ;
- `Sexe` ;

Trois variables `Z1`, `Z2`, `Z3` qui décrivent les caractéristiques du sujet :

- `Z1` = variable qualitative nominale à trois modalités :
  - sain,
  - pathologique,
  - contrôle ;
- `Z2` = variable qualitative nominale à trois modalités :
  - amateur,
  - professionnel,
  - contrôle ;
- `Z3` = variable qualitative nominale à trois modalités :
  - soliste,
  - choriste,
  - contrôle ;

Deux variables `F1` et `F2` qui décrivent la sévérité de l'aspect fonctionnel du trouble :
- `F1` = score obtenu lors de la première évaluation,
- `F2` = score obtenu lors de la seconde évaluation ;

4. Le tableau de données se trouve en annexe 2.

Deux variables E1 et E2 qui décrivent la sévérité de l'aspect émotionnel du trouble :

- E1 = score obtenu par le sujet lors de la première évaluation,
- E2 = score obtenu par le sujet lors de la seconde évaluation ;

Deux variables P1 et P2 qui décrivent la sévérité de l'aspect physique du trouble :

- P1 = score obtenu par le sujet lors de la première évaluation,
- P2 = score obtenu par le sujet lors de la seconde évaluation ;

Deux variables VHI1 et VHI2 qui reflètent l'impact général du trouble vocal sur la vie quotidienne du sujet :

- VHI1 = résultat global obtenu par le sujet lors de la première évaluation,
- VHI2 = résultat global obtenu par le sujet lors de la seconde évaluation.

Pour pouvoir évaluer l'efficacité du questionnaire du VHI pour les chanteurs, Simon a comparé les résultats du test des chanteurs pathologiques, des chanteurs sains et des non chanteurs.

Elle a utilisé le test non paramétrique de Wilcoxon pour des échantillons indépendants pour comparer les scores :

- des chanteurs sains à ceux des chanteurs pathologiques ;
- des chanteurs professionnels à ceux des chanteurs amateurs ;
- des chanteurs solistes à ceux des chanteurs choristes.

Ses analyses statistiques révèlent que :

- les résultats généraux des chanteurs pathologiques sont nettement plus élevés que ceux des chanteurs sains ;
- les résultats généraux des chanteurs professionnels sont inférieurs à ceux des chanteurs amateurs. Cette tendance est également retrouvée au niveau des questions traitant des aspects fonctionnel et physique du trouble ;
- les choristes obtiennent en moyenne des résultats plus élevés que les solistes. Les différences sont significatives pour les questions traitant des aspects fonctionnel et physique ;
- le fait d'être professionnel a tendance à faire chuter le score total, tandis que le fait d'être pathologique le fait plutôt augmenter ;
- le sexe n'influence nullement les résultats du test.

Simon conclu en disant que les résultats des analyses statistiques sont tout à fait satisfaisants. Le fait que les résultats des chanteurs pathologiques sont nettement plus élevés que ceux des chanteurs sains indique que les chanteurs pathologiques peuvent évaluer leur trouble vocal au travers des questions et qu'ils peuvent coter le niveau de sévérité de leur trouble. En effet, un score plus élevé reflète un degré de handicap perçu plus élevé.

5.1.4 Motivation de l'étude

Pour son étude, Simon demande aux sujets de remplir le questionnaire du VHI adapté à la voix chantée à deux reprises. Bien qu'aucun traitement n'a été appliqué entre les deux évaluations, il apparaît que les scores obtenus lors de la seconde évaluation sont significativement inférieurs à ceux obtenus lors de la première évaluation.

Lorsque le problème nous a été présenté, nous avons proposé de reprendre le jeu de données de Simon et de le transformer en un ensemble de données symboliques afin de prendre en compte la variabilité existante entre les deux évaluations.

En appliquant des méthodes de classification symboliques sur les données transformées, nous espérons pouvoir extraire un maximum d'information. Par définition, la classification cherche à regrouper des sujets avec des comportements semblables et de différencier autant que possible des sujets avec des comportements différents.

5.1.5 Données symboliques

Nous avons repris les données récoltées par Simon lors de son étude et nous les avons transformées en un ensemble de données symboliques. C'est à partir du tableau de données symboliques[5] que nous allons travailler.

Les variables symboliques sont les suivantes :

- `Age` : Variable univaluée quantitative ;
- `Sexe` : Variable univaluée catégorique à deux modalités :
  - homme,
  - femme ;
- `Z1` : Variable univaluée catégorique à trois modalités :
  - sain,
  - pathologique,
  - contrôle ;

5. Le tableau de données symboliques se trouve en annexe 3.

- Z2 : Variable univaluée catégorique à trois modalités :
  - amateur,
  - professionnel,
  - contrôle ;
- Z3 : Variable univaluée catégorique à trois modalités :
  - soliste,
  - choriste,
  - contrôle ;
- VHI : Variable intervalle.

La variable VHI mesure l'impact global du trouble vocal sur la vie quotidienne du sujet. Les bornes inférieure et supérieure de l'intervalle correspondent respectivement au plus petit et au plus grand des deux scores obtenus par le sujet lors des évaluations.

Remarque : Nous n'avons pas considéré les variables traitant des aspects fonctionnel, émotionnel et physique du trouble vocal dans notre analyse parce qu'il semble y avoir eu quelques confusions au niveau des scores obtenus par certains patients pour les questions concernant l'aspect physique du trouble.

5.1.6 Classification des sujets

Pour notre étude, nous avons opté pour la méthode de classification symbolique SCLUST parce qu'elle permet le traitement simultané de données de types différents.

Lors de la paramétrisation de l'algorithme, nous avons choisi de ne pas normaliser les données. Ainsi, toutes les variables ont le même poids dans l'analyse.

Pour décider du nombre de classes à retenir, nous nous sommes basés sur l'analyse de Simon. En appliquant le test de Wilcoxon pour comparer les scores entre les différents groupes de sujets (chanteurs pathologiques vs chanteurs sains, amateurs vs professionnels, choristes vs solistes), Simon arrivait souvent à la conclusion que les résultats d'un groupe étaient significativement supérieurs ou inférieurs à ceux de l'autre groupe considéré.

En retenant un petit nombre de classes, nous pourrons déterminer si la méthode SCLUST parvient à identifier les éventuelles différences entre certains groupes de sujets.

Nous considérons la partition de l'ensemble des sujets en 4 classes.

Nous avons réinitialisé 50 fois l'algorithme. La partition en quatre classes produite par SCLUST est décrite dans ce qui suit.

La répartition des sujets dans les quatre classes est reprise dans le tableau 5.1 :

Classe	Effectifs de la classe
1	51
2	35
3	44
4	22

TAB. 5.1 – *Composition des classes.*

Nous constatons que la classe 1 compte 51 sujets soit le tiers de l'ensemble des sujets interrogés. La classe 4 en revanche ne compte que très peu de sujets (22).

En assimilant les sujets aux prototypes des classes les plus proches, nous remarquons que 75% de l'inertie totale est expliquée.

Le tableau 5.8 nous informe sur la contribution des variables à la formation de la partition :

Variable	Pouvoir discriminant par rapport à la partition	Contribution relative à l'inertie intra-classe	Contribution relative à l'inertie totale
Age	45.23	12.12	20.21
Sexe	0.67	0.00	0.04
Z1	25.88	0.02	0.06
Z2	3.74	0.00	0.05
Z3	6.70	0.01	0.06
VHI	83.20	87.85	79.58

TAB. 5.2 – *Contribution de chaque variable à la formation de la partition.*

Au vu de ce tableau, nous constatons que la variable la plus discriminante est la variable VHI. En moindre mesure, nous avons également les variables Age et Z1.

En regardant le tableau, nous pouvons d'ores et déjà dire que ce sont principalement les résultats des sujets au test, leur âge et leur "santé" vocale qui font que les sujets se trouvent dans une classe plutôt que dans une autre.

Pour pouvoir identifier les principales différences entre les sujets des classes, nous allons analyser les prototypes des classes.

Le tableau 5.3 nous informe sur l'âge moyen des sujets pour chaque classe :

Classe	Age moyen des sujets présents dans la classe
1	39
2	29
3	54
4	46

TAB. 5.3 – *Age moyen des sujets de chaque classe.*

La moyenne d'âge de l'ensemble des sujets interrogés est de 42 ans. Sachant cela, nous constatons que la moyenne d'âge de la classe 3 est nettement supérieure à celle de la population interrogée. Ce qui n'est pas le cas pour la classe 2. La moyenne d'âge des sujets de cette classe est de 29 ans.

Le tableau 5.4 décrit la répartition des hommes et des femmes au sein de chaque classe :

Modalité	Population	Classe 1	Classe 2	Classe 3	Classe 4
Homme	0.25	0.25	0.20	0.30	0.23
Femme	0.75	0.75	0.80	0.70	0.77

TAB. 5.4 – *Fréquence d'hommes et de femmes dans chaque classe.*

Dans la population étudiée, 75% des sujets sont des femmes, contre seulement 25% de sujets de sexe masculin. En parcourant le tableau, nous constatons que cette proportion d'hommes et de femmes est plus ou moins retrouvée dans chaque classe. Ceci nous pousse à croire que le sexe des sujets n'influence pas le fait qu'ils se trouvent dans une classe plutôt que dans une autre.

Le tableau 5.5 décrit la répartition des chanteurs pathologiques, des chanteurs sains et des non chanteurs dans les classes :

Modalité	Population	Classe 1	Classe 2	Classe 3	Classe 4
Chanteur sain	0.63	0.55	0.77	0.86	0.09
Chanteur pathologique	0.24	0.31	0.03	0.05	0.82
Sujet contrôle	0.13	0.14	0.20	0.09	0.09

TAB. 5.5 – *Fréquence de chanteurs sains et pathologiques dans chaque classe.*

Tout d'abord, nous remarquons que :
- 63% des sujets interrogés sont des chanteurs sains ;
- 24% des enquêtés sont des chanteurs pathologiques ;
- 13% des sujets interrogés ne sont pas chanteurs.

En analysant le tableau, nous constatons au niveau des classes que :
- 86% des sujets de la classe 3 sont des chanteurs sains ;
- 82% des sujets de la classe 4 sont des chanteurs pathologiques, c'est-à-dire des chanteurs qui se plaignent d'un quelconque trouble vocal ;
- la classe 2 contient 77% de chanteurs sains et 20% de sujets contrôles. Cette classe ne compte pratiquement pas de chanteurs pathologiques.

Le tableau 5.6 nous indique la manière dont l'ensemble des chanteurs amateurs et des chanteurs professionnels est divisé :

Modalité	Population	Classe 1	Classe 2	Classe 3	Classe 4
Chanteur amateur	0.71	0.68	0.57	0.73	0.91
Chanteur professionnel	0.16	0.18	0.23	0.18	0.00
Sujet contrôle	0.13	0.14	0.20	0.09	0.09

TAB. 5.6 – *Fréquence d'amateurs et de professionnels dans chaque classe.*

Des sujets interrogés, nous voyons que :
- 71% sont des chanteurs amateurs ;
- 16% sont des chanteurs professionnels ;
- 13% ne sont pas chanteurs.

Au vu de ce tableau, nous remarquons que la classe 4 est constituée uniquement de chanteurs amateurs. Autrement dit, il n'y a pas de chanteurs professionnels dans cette classe. La répartition des sujets de la population est plus ou moins retrouvée au sein des trois autres classes, excepté peut-être pour la classe 2 où le pourcentage de chanteurs professionnels est un peu plus important.

Le tableau 5.7 détaille la répartition des chanteurs solistes et des chanteurs choristes au sein des quatre classes :

Modalité	Population	Classe 1	Classe 2	Classe 3	Classe 4
Chanteur soliste	0.61	0.67	0.74	0.50	0.45
Chanteur choriste	0.26	0.20	0.06	0.41	0.45
Sujet contrôle	0.13	0.14	0.20	0.09	0.09

TAB. 5.7 – *Fréquences des chanteurs solistes et des choristes dans chaque classe.*

Au vu de ce tableau, nous constatons tout d'abord que :
- 61% des sujets interrogés sont des solistes ;
- 26% des enquêtés sont des choristes ;
- 13% des sujets interrogés ne sont pas chanteurs.

Au niveau des classes, nous remarquons que :
- la classe 2 est composée principalement de chanteurs solistes ;
- les classes 3 et 4 comptent plus ou moins le même pourcentage de chanteurs solistes que de chanteurs choristes.

Les prototypes des classes suivant la variable VHI sont représentés à la figure 5.1 :

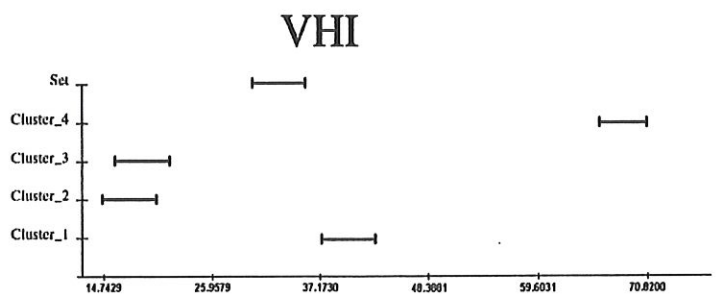

FIG. 5.1 – *Prototypes des classes pour la variable VHI.*

Nous constatons que les prototypes des classes sont assez bien séparés selon cette variable. La variable VHI permet de discriminer les classes 2 et 3 de la classe 1 et de la classe 4.

Le prototype de la classe 4 est particulièrement élevé. La position du prototype illustre le fait que les sujets de cette classe sont relativement affectés par leur trouble vocal. En moyenne, leur score se situe entre 66 et 72.

Les prototypes des classes 2 et 3 se trouvent à l'extrême gauche sur la figure 5.1. Le score des sujets de ces classes est relativement faible. Il varie entre 15 et 20 pour l'une, entre 16 et 22 pour l'autre.

Le prototype de la classe 1 se trouve à proximité du prototype de la population. Les sujets de la classe 1 obtiennent en moyenne un score compris entre 37 et 43. Le score moyen de l'ensemble des 152 enquêtés se situe quant à lui entre 30 et 35.

En résumé, nous pouvons dire que :

- la classe 1 compte 51 sujets soit le tiers de la population interrogée. Les sujets de cette classe sont ceux qui présentent un comportement moyen ;
- les sujets de la classe 2 sont des chanteurs sains solistes relativement jeunes ;
- les sujets de la classe 3 sont plus âgés. Ils ont en moyenne 54 ans. Ce sont principalement des chanteurs amateurs sains. Ce qui justifie probablement leur faible score au test ;
- les sujets de la classe 4 ont en moyenne 46 ans. Les sujets de la classe sont des chanteurs amateurs pathologiques. En analysant le prototype de la classe selon la variable `VHI`, nous constatons que ces chanteurs sont particulièrement affectés par leur trouble vocal.

Dans la partition optimale trouvée par SCLUST, les sujets sont répartis de la manière suivante :

- les jeunes chanteurs solistes sains ;
- les chanteurs amateurs sains plus âgés ;
- les chanteurs amateurs pathologiques.

Finalement, nous pouvons dire que nous retrouvons en quelque sorte les résultats obtenus par Simon. En effet, le fait que SCLUST sépare les sujets de la sorte nous pousse à croire qu'il y a des différences assez importantes au niveau des scores obtenus au VHI entre les chanteurs pathologiques et les chanteurs sains. Sur base de ces résultats, nous pouvons conclure que le questionnaire du *voice handicap index* est bien adapté à la voix chantée.

5.2 Application 2 : Union Européenne

Dans le cadre de cette étude, nous nous intéressons aux pays membres de l'Union Européenne. Le 1^{er} mai 2004, dix nouveaux pays ont adhéré à l'Union Européenne, élargissant ainsi à nouveau ses frontières, et portant le nombre de ses membres à 25. Parmi ces pays figurent Chypre, l'Estonie, la Hongrie, la Lettonie, la Lituanie, Malte, la Pologne, la République Slovaque, la République Tchèque et la Slovénie.

Pendant de nombreuses années, l'Europe a été divisée en deux parties : l'Europe de l'Ouest et l'Europe de l'Est. L'objectif de notre étude est de déterminer si cette division de l'Europe est encore perceptible aujourd'hui.

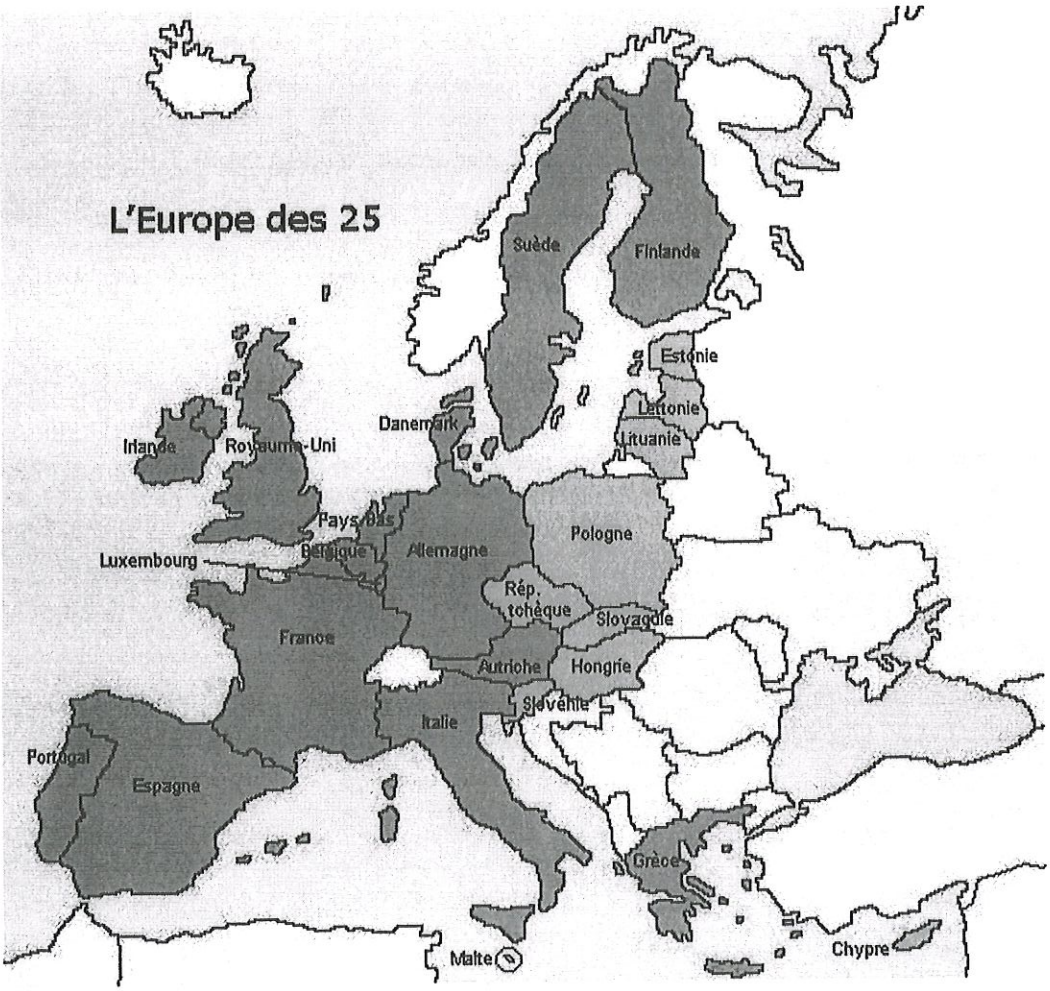

FIG. 5.2 – *Carte de l'Europe des 25.*

5.2.1 Le jeu de données

Les données utilisées pour cette étude proviennent du site de la Banque Mondiale.

Chaque pays de l'Union Européenne est ainsi décrit par un certain nombre d'indicateurs de développement, mesurés chaque année entre 1999 et 2003.

Les variables mesurées sur les différents pays sont les suivantes :

- les indicateurs économiques :
  - `PIB par habitant` = la valeur par habitant de tous les biens et services finaux produits sur le territoire du pays considéré durant une année. C'est une façon de mesurer les richesses créées dans un pays (en USD),
  - `Taux de croissance du PIB` (en %),
  - `Exportation des biens et services` représente l'ensemble des biens et services fournis par des résidents à des non-résidents (en % du PIB),
  - `Importation des biens et services` représente l'ensemble des biens et des services fournis par des non-résidents à des résidents (en % du PIB),
  - `Taux d'inflation` = taux d'accroissement du niveau des prix (en %) ;
- les indicateurs démographiques :
  - `Taux de croissance de la population` (en %),
  - `Taux de fertilité` = rapport du nombre de naissances vivantes au nombre de femmes en âge de procréer (en nombre d'enfants par femme),
  - `Taux de mortalité infantile` = rapport entre le nombre d'enfants décédés à moins d'un an et l'ensemble des enfants nés vivants (en ‰),
  - `Espérance de vie à la naissance` = la durée moyenne de la vie de la naissance au décès d'une population (en années) ;
- au niveau de l'éducation :
  - `Pourcentage d'inscriptions aux études secondaires` = rapport du nombre d'enfants inscrits aux études secondaires au nombre d'enfants en âge d'être à ce niveau d'études (en %) ;
- les indicateurs de consommation :
  - `Consommation d'électricité` (en kWh par habitant),
  - `Consommation d'énergie` (en kg de carburant par habitant),
  - `Émission de CO2` (en tonnes par habitant) ;

- au niveau de la communication :
  - `Nombre de souscripteurs de lignes téléphoniques fixes et téléphones mobiles` (par 1000 habitants),
  - `Nombre d'ordinateurs personnels` (par 1000 habitants),
  - `Nombre d'utilisateurs d'Internet` (par 1000 habitants).

Les variables présentées ci-dessus ont été mesurées chaque année entre 1999 et 2003 pour chaque pays. Nous disposons donc, par pays, de cinq mesures pour chaque variable. Ces différentes mesures décrivent l'évolution des indicateurs de développement entre 1999 et 2003.

Pour tenir compte cette évolution des indicateurs, nous allons travailler avec des données symboliques de type intervalle. Les bornes inférieure et supérieure d'un intervalle seront respectivement les valeurs minimale et maximale prise par la variable considérée pour un pays donné au cours des cinq années de notre historique[6].

Nous débutons notre étude par une classification des pays de l'Union Européenne sur base de l'entièreté des indicateurs de développement. Notre objectif étant de déterminer si les méthodes de classification symboliques isolent les pays de l'Europe de l'Est[7] des autres pays de l'Union Européenne, nous effectuerons uniquement une partition des pays en deux classes.

6. Le tableau de données symboliques se trouve en annexe 4.

7. Nous entendons par pays de l'Europe de l'Est : l'Estonie, la Hongrie, la Lettonie, la Lituanie, la Pologne, la République Slovaque, la République Tchèque et la Slovénie.

5.2.2 Classification des pays sur base de toutes les variables

5.2.2.1 La méthode SCLUST

Étant donné qu'il y a des différences au niveau des ordres de grandeurs dans les données, nous avons choisi, lors de la paramétrisation de la méthode, de normaliser les données. Ainsi, chaque variable à la même importance quelle que soit sa dispersion.

Nous avons réinitialisé 50 fois l'algorithme. La partition optimale trouvée par SCLUST après quelques itérations seulement est présentée ci-dessous :

```
Classe :   1  Cardinal :    13
================================

(  0)  Allemagne      [0.5]   (  1)  Autriche      [0.3]   (  2)  Belgique      [0.5]
(  3)  Chypre         [0.7]   (  4)  Danemark     [0.4]   (  7)  Finlande      [0.9]
(  8)  France         [0.5]   ( 11)  Irlande      [2.0]   ( 12)  Italie        [0.9]
( 15)  Luxembourg     [4.5]   ( 17)  Pays-Bas     [0.3]   ( 22)  Royaume-Uni   [0.4]
( 24)  Suede          [1.3]

Classe :   2  Cardinal :    12
================================

(  5)  Espagne        [1.1]   (  6)  Estonie      [1.2]   (  9)  Grece         [1.0]
( 10)  Hongrie        [1.3]   ( 13)  Lettonie     [1.6]   ( 14)  Lituanie      [1.4]
( 16)  Malte          [1.6]   ( 18)  Pologne      [0.6]   ( 19)  Portugal      [0.7]
( 20)  Rep.Slovaque   [0.3]   ( 21)  Rep.Tcheque  [0.6]   ( 23)  Slovenie      [0.7]
```

Nous constatons que l'algorithme de partitionnement sépare l'Europe en deux groupes comptant respectivement 13 et 12 pays.

Dans la classe 2, nous retrouvons tous les pays de l'Europe de l'Est (Estonie, Hongrie, Lettonie, Lituanie, Pologne, République Slovaque, République Tchèque et Slovénie), ainsi que l'Espagne, la Grèce, Malte et le Portugal.

Les valeurs entre crochets représentent la distance entre un pays et le prototype de la classe à laquelle il appartient. Si nous regardons ces valeurs, nous remarquons que le Luxembourg, l'Irlande et la Suède s'écartent assez bien du prototype de la classe 1. Au plus un pays est proche du prototype de la classe à laquelle il appartient, au plus il y a des ressemblances entre ce pays et le prototype au niveau de leurs caractéristiques.

Caractérisation des classes

Le tableau 5.8 nous informe sur la contribution des variables à la formation de la partition :

Liste de variables	Pouvoir discriminant par rapport à la partition	Contribution relative à l'inertie intra-classe
PIB par habitant	72.64	14.75
Taux de croissance du PIB	1.94	0.39
Exportations des biens et services	0.24	0.05
Importations des biens et services	2.21	0.45
Taux d'inflation	23.87	4.85
Taux de croissance de la population	15.07	3.06
Taux de fertilité	52.22	10.61
Taux de mortalité infantile	28.83	5.86
Espérance de vie à la naissance	43.74	8.88
Inscriptions aux études secondaires	7.23	1.47
Consommation d'électricité	35.87	7.29
Consommation d'énergie	40.71	8.27
Émission de CO2	12.91	2.62
Lignes téléphoniques et portables	51.94	10.55
Nbre d'ordinateurs personnels	57.98	11.78
Nbre d'utilisateurs d'Internet	44.93	9.12

TAB. 5.8 – *Contribution de chaque variable à la formation de la partition.*

Au vu de ce tableau, nous constatons que la variable la plus discriminante est la variable PIB par habitant, viennent ensuite les variables Nbre d'ordinateurs personnels , Taux de fertilité et Lignes téléphoniques et portables. La contribution relative de ces variables à l'inertie intra-classe est nettement plus élevée que pour les autres variables.

Dans le but de pouvoir identifier les principales différences entre les pays des deux classes, nous avons représenté les prototypes des classes.

Les prototypes des classes suivant la variable PIB par habitant sont présentés à la figure 5.3.

Sur cette figure, nous constatons que la variable PIB par habitant permet de séparer les pays de l'Union Européenne en deux classes. Si nous regardons la position des prototypes sur le graphique, nous constatons que l'écart entre les deux est énorme. En moyenne, le PIB par habitant varie entre 24.129 et 27.223 USD pour les pays de la classe 1, entre 6.931 et 8.564 USD seulement pour les pays de la classe 2.

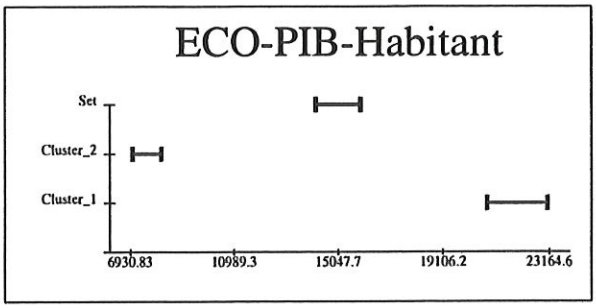

FIG. 5.3 – *Prototypes des classes pour la variable* `PIB par habitant`.

Si nous regardons le tableau de données, nous remarquons que le PIB par habitant est très bas pour certains pays de la classe 2. Pour les pays baltes[8] par exemple, le PIB par habitant varie entre 2.810 et 5.380 USD.

Les prototypes des classes suivant la variable `Nbre d'ordinateurs personnels` sont représentés à la figure 5.4 :

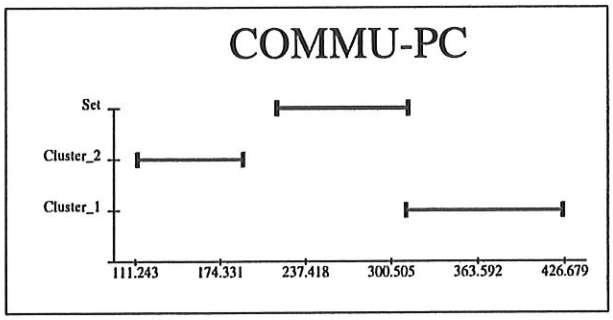

FIG. 5.4 – *Prototypes des classes pour la variable* `Nbre d'ordinateurs personnels` *pour 1000* `habitants`.

En moyenne, le nombre d'ordinateurs personnels pour 1000 habitants est nettement plus élevé pour les pays de la classe 1. Il varie entre 310 et 427 ordinateurs pour 1000 habitants pour la classe 1, alors qu'il se situe entre 111 et 193 pour la classe 2.

8. L'Estonie, la Lettonie et la Lituanie.

Les prototypes des classes suivant la variable `Taux de fertilité` sont représentés à la figure 5.5 :

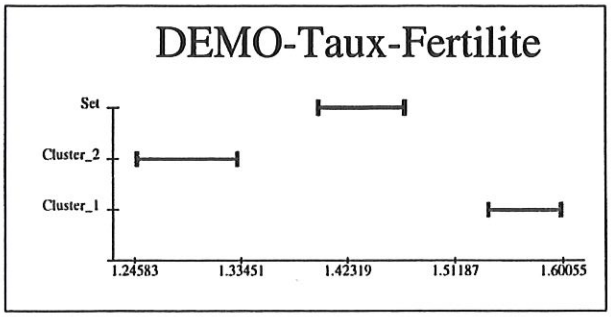

FIG. 5.5 – *Prototypes des classes pour la variable `Taux de fertilité`.*

En regardant le graphique, nous constatons que le taux de fertilité des pays de la classe 1 est en moyenne légèrement supérieur à celui des pays de la classe 2.

Les prototypes des classes suivant la variable `Souscripteurs de lignes fixes et téléphones mobiles` sont représentés à la figure 5.6 :

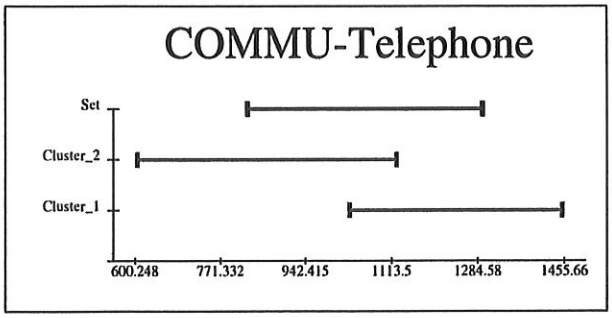

FIG. 5.6 – *Prototypes des deux classes pour la variable `Souscripteurs de lignes fixes et téléphones mobiles`.*

Nous constatons sur le graphique que les prototypes des deux classes se recoupent suivant cette variable. La longueur des intervalles des prototypes indique qu'il y a une grande variabilité entre les pays d'une même classe en ce qui concerne cette variable. D'ailleurs, la distinction entre les pays de l'Union Européenne est moins nette suivant cette variable. Dans l'ensemble, nous pouvons tout de même dire qu'il y a, en moyenne, plus de souscripteurs de lignes fixes et de téléphones portables dans les pays de la classe 1.

5.2.2.2 Méthode Div

La hiérarchie de partitions produite par la méthode Div est la suivante :

```
PARTITION IN 2 CLUSTERS  :
-------------------------

Cluster 1 (n=13) :  Chypre  Espagne  Estonie   Grece  Hongrie  Lettonie   Lituanie
                    Malte   Pologne  Portugal  Rep.Slovaque  Rep.Tcheque  Slovenie

Cluster 2 (n=12) :  Allemagne  Autriche  Belgique    Danemark  Finlande  France
                    Irlande    Italie    Luxembourg  Pays-Bas  Royaume-Uni  Suede

Explicated inertia : 74.640450
```

Lors de la paramétrisation de la méthode, nous avons ici aussi choisi de normaliser les données. Cette normalisation s'est faite par l'inverse de la dispersion.

La description de la partition en deux classes ci-dessous illustre la manière dont les pays de l'Union Européenne sont classés :

```
Cluster 1 :
   IF    1- [ECO_PIB_Habitant <= 18052.500000] IS TRUE

Cluster 2 :
   IF    1- [ECO_PIB_Habitant <= 18052.500000] IS FALSE
```

Nous constatons que la coupure de l'ensemble des pays de l'Union Européenne en deux classes se fait suivant la variable `PIB par habitant`. Ceci souligne une fois de plus le pouvoir discriminant de cette variable.

En répartissant les pays de la sorte, nous constatons qu'une part considérable de l'inertie est expliquée. La partition en deux classes représentée ci-dessus permet d'expliquer près de 75% de l'inertie contenue dans les données.

Si nous comparons cette partition en deux classes avec celle trouvée par Sclust, nous sommes tentés de dire qu'elles sont identiques. À vrai dire, seul Chypre a changé de classe. Pour la partition trouvée par Sclust, Chypre se trouve avec les pays de l'Europe de l'Ouest. L'algorithme Div classe ce pays avec les pays de l'Europe de l'Est.

Si nous consultons le tableau de données, nous voyons que la valeur du PIB par habitant pour Chypre varie entre 12.220 et 12.460 USD. Le PIB par habitant est donc inférieur à la valeur de coupure 18.052 USD. Ceci explique la raison pour laquelle Chypre a changé de classe.

L'observation concernant Chypre est plutôt intéressante dans la mesure où elle nous révèle que si nous considérons uniquement la variable PIB par habitant pour séparer les pays de l'Union Européenne, Chypre se trouve dans la même classe que les pays de l'Europe de l'Est, alors que si nous considérons l'ensemble des variables pour la séparation des pays, Chypre se trouve dans la même classe que les pays de l'Ouest.

L'arbre de classification obtenu par la méthode DIV est représenté à la figure 5.7 :

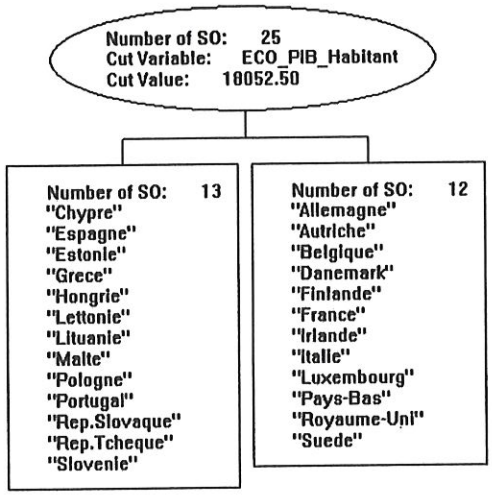

FIG. 5.7 – *Arbre de classification.*

5.2.2.3 La méthode Sclass

Les résultats obtenus avec la méthode SCLASS sont détaillés ci-dessous :

```
Split of the node :     1
=========================

Number of Symbolic objects in the node :      25pt
---------------------------------------

Criteria of cut :
-----------------

Cut variable :( 16) COMMU_Internet
Cut value : 219.31
Smoothing parameter CENTER :  44.37
Smoothing parameter LENGTH :  12.41

Rule : if value of i <  219.31 -> the SO i is in the left node
       if value of i >  219.31 -> the SO i is in the right node
```

La coupure du noeud 1 se fait suivant la variable Nombre d'utilisateurs d'Internet par 1000 habitants.

La division du noeud 1 engendre les deux noeuds suivants :

```
Node :   2    Cardinal :  10pt
================================
 (5)  Espagne
 (9)  Grece
(10)  Hongrie
(11)  Irlande
(14)  Lituanie
(16)  Malte
(18)  Pologne
(19)  Portugal
(20)  Rep.Slovaque
(21)  Rep.Tcheque

Node :   3    Cardinal :  15pt
================================
 (0)  Allemagne
 (1)  Autriche
 (2)  Belgique
 (3)  Chypre
 (4)  Danemark
 (6)  Estonie
 (7)  Finlande
 (8)  France
(12)  Italie
(13)  Lettonie
(15)  Luxembourg
(17)  Pays-Bas
(22)  Royaume-Uni
(23)  Slovenie
(24)  Suede
```

En comparant la composition de ces deux noeuds avec la partition en deux classes produite par SCLUST, nous constatons que :

- l'Estonie, la Lettonie et la Slovénie sont classés avec les pays de l'Europe de l'Ouest (noeud 3) ;
- l'Irlande ne fait plus partie de la classe des pays de l'Europe de l'Ouest (noeud 2).

Une fois le processus de division terminé, nous obtenons l'arbre de classification représenté à la figure 5.8.

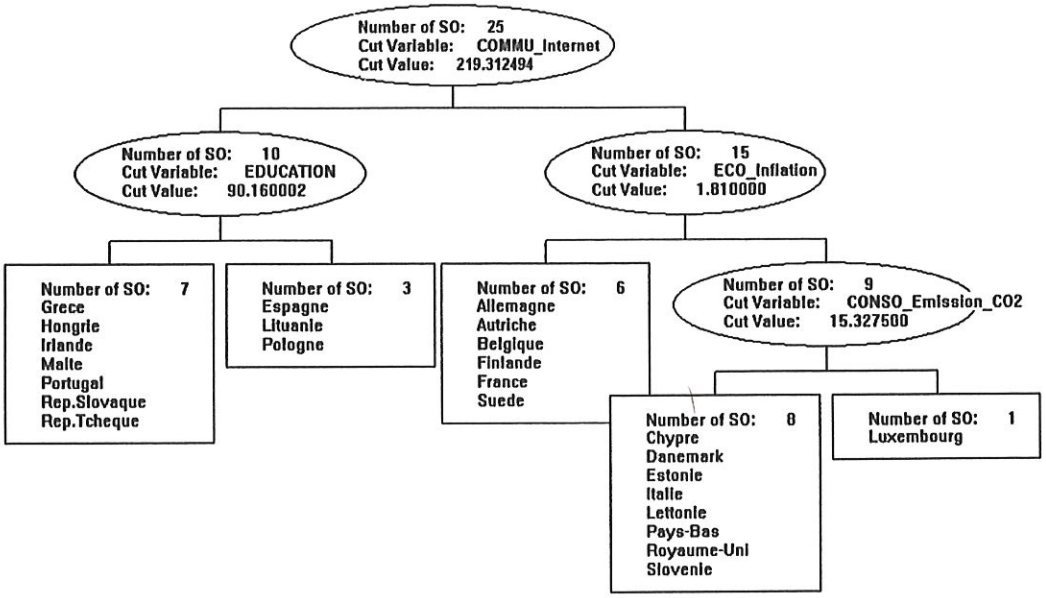

FIG. 5.8 – *Arbre de classification.*

5.2.2.4 La méthode HIPYR

La hiérarchie de partitions produite par la méthode HIPYR avec le critère du degré de généralité est représentée à la figure 5.9 :

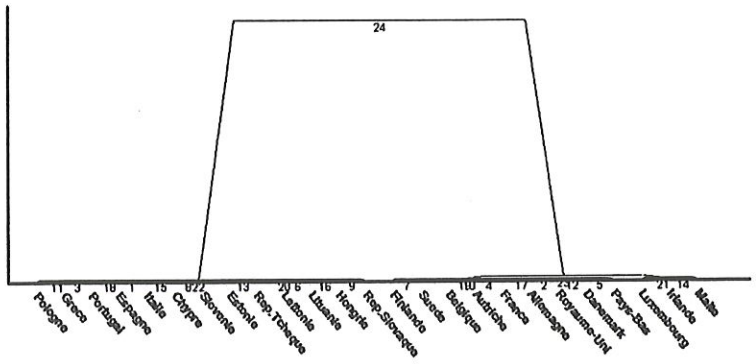

FIG. 5.9 – *Hiérarchie de partitions - Critère du degré de généralité.*

Comme nous pouvons le constater, cette partition en deux classes est quasiment identique à celle trouvée par SCLUST. Seuls l'Italie, Chypre et Malte ont changé de classe. Ici, l'Italie et Chypre se retrouvent dans la classe composée de l'Espagne, du Portugal, des pays baltes, etc, tandis que Malte se retrouve dans la classe composée de la plupart des pays de l'Ouest.

En regardant le dendrogramme, nous constatons que le niveau d'agrégation des deux dernières classes est extrêmement élevé. Cela signifie qu'il y a de nettes différences de caractéristiques entre les deux classes.

5.2.2.5 Conclusion

Nous ne connaissons pas la "meilleure" partition en deux classes des 25 pays membres de l'Union Européenne. Nous ne pouvons donc pas dire laquelle des partitions obtenues ci-dessus est la bonne. Par contre, après avoir analysé les différentes partitions en deux classes trouvées par chacune des méthodes, nous pouvons dire qu'elles se ressemblent assez bien.

Au cours de notre analyse, nous avons constaté que :

- l'Allemagne, l'Autriche, la Belgique, le Danemark, la Finlande, la France, le Luxembourg, les Pays-Bas, le Royaume-Uni et la Suède sont toujours classés ensemble au sein d'une classe ;
- l'Espagne, la Grèce, la Hongrie, la Lituanie, la Pologne, le Portugal, la République Slovaque et la République Tchèque sont toujours classés ensemble au sein d'une autre classe.

Le fait que certains pays se retrouvent toujours ensemble nous encourage à croire qu'il y a des ressemblances assez fortes entre eux.

Sur base des variables de développement présentes dans le tableau de données, nous pouvons dire que certains pays de l'Europe de l'Ouest tels l'Allemagne, la France, la Belgique sont des pays qui se ressemblent assez bien. Ils ont des caractéristiques et des tendances similaires. Cette remarque vaut aussi pour les pays de l'Europe de l'Est.

En séparant les pays de l'Union Européenne en deux groupes, nous retrouvons quelque part cette division Est-Ouest de l'Europe. Une grande différence entre les deux parties de l'Europe se situe au niveau de la valeur du PIB par habitant. Celle-ci est relativement faible pour les pays de l'Est.

5.2.3 Classification des pays sur base des indicateurs économiques

Pour affiner notre étude, nous effectuons dans cette section une classification des pays uniquement à partir des indicateurs économiques. Nous espérons pouvoir identifier les tendances économiques existantes au sein de l'Union Européenne.

Étant donné que les méthodes de classification produisent pratiquement toutes la même partition, nous nous contentons par la suite d'analyser uniquement la sortie produite par SCLUST.

5.2.3.1 Résultats de la méthode SCLUST

Cette fois encore, nous avons normalisé les données. Nous donnons ainsi la même importance à chacune des variables quelle que soit sa dispersion.

Le partition optimale trouvée par SCLUST est présentée ci-dessous :

```
Classe :   1  Cardinal :      7
================================

(  6) Estonie          [0.3]  ( 10)  Hongrie       [1.1]  ( 11) Irlande          [0.9]
( 15) Luxembourg       [2.9]  ( 16)  Malte         [0.5]  ( 20) Rep.Slovaque     [0.5]
( 23) Slovenie         [0.9]

Classe :   2  Cardinal :     18
================================

(  0) Allemagne        [0.7]  (  1) Autriche       [0.5]  (  2) Belgique         [1.9]
(  3) Chypre           [0.4]  (  4) Danemark       [0.9]  (  5) Espagne          [0.4]
(  7) Finlande         [0.4]  (  8) France         [0.5]  (  9) Grece            [1.0]
( 13) Lettonie         [2.5]  ( 14) Lituanie       [3.0]  ( 17) Pays-Bas         [1.0]
( 18) Pologne          [1.4]  ( 19) Portugal       [0.8]  ( 21) Rep.Tcheque      [1.2]
( 22) Royaume-Uni      [0.6]  ( 24) Suede          [0.4]
```

Cette partition est assez différente de celle trouvée avec toutes les variables. Cette partition peut d'ailleurs paraître assez étonnante a priori dans la mesure où nous retrouvons le Luxembourg avec des pays de l'Est tels que l'Estonie, la Hongrie, la République Slovaque et la Slovénie.

Le tableau 5.9 reprend la contribution relative des variables à la formation de la partition.

Liste de variables	Pouvoir discriminant par rapport à la partition	Contribution relative à l'inertie intra-classe
PIB par habitant	1.90	1.11
Taux de croissance du PIB	21.18	12.41
Exportations des biens et services	53.59	31.39
Importations des biens et services	57.76	33.84
Taux d'inflation	36.28	21.25

TAB. 5.9 – *Contribution de chaque variable.*

En regardant ce tableau, nous constatons que la variable PIB par habitant n'est pas la variable la plus discriminante parmi les indicateurs de développement économiques. Les variables les plus discrimantes ici sont les variables Importations des biens et services et Exportations des biens et services. En moindre mesure, nous avons également les variables Taux d'inflation et Taux de croissance du PIB.

Si nous regardons le tableau de données de départ, nous constatons que les tendances des pays de la classe 1 sont assez différentes de celles des pays de la classe 2. Pour les pays de la classe 1, nous observons que :

- le pourcentage d'exportation et d'importation de biens et de services est relativement important. L'exportation et l'importation représente plus de 60% du PIB ;
- le taux d'inflation est élevé. Pour la plupart pays, il atteint 5%.

La Belgique, les Pays-Bas et République Tchèque présentent des caractéristiques similaires pour les variables liées à l'exportation et l'importation des biens et services. Par contre, le taux d'accroissement du PIB de ces trois pays est nettement plus faible que celui des pays de la classe 1.

5.2.3.2 Caractérisation des classes

Nous allons à présent tenter de caractériser les classes de manière à avoir une idée des caractéristiques générales des pays qui les composent.

Les prototypes des classes pour les variables Exportations des biens et services et Importations des biens et services sont représentés aux figures 5.10 et 5.11.

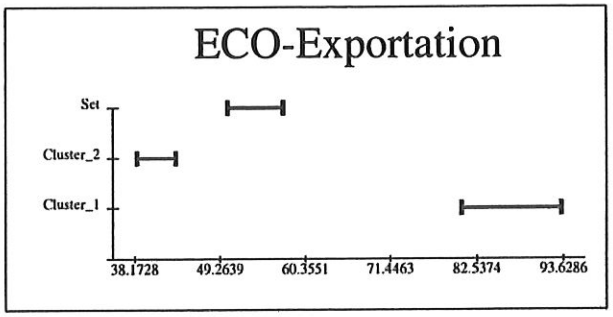

FIG. 5.10 – *Prototypes des classes pour la variable* `Exportations des biens et services.`

Nous constatons sur ce graphique que les prototypes des classes sont bien séparés suivant cette variable.

Le prototype de la classe 1 est assez élevé. Les bornes inférieure et supérieure de l'intervalle sont respectivement 80 et 94. Ceci signifie que l'exportation des biens représente une part considérable du PIB. Le prototype de la classe 2 se trouve à l'extrême gauche sur le graphique. L'exportation des biens et services est donc nettement moins importante pour les pays de cette classe.

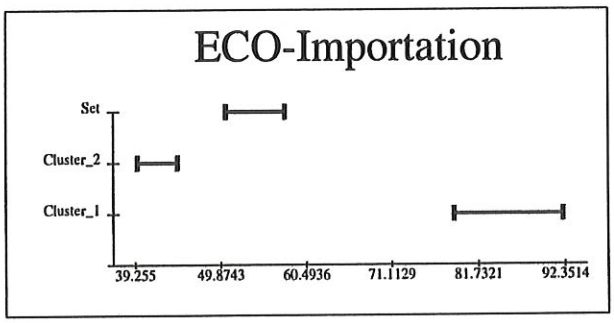

FIG. 5.11 – *Prototypes des classes pour la variable* `Importations des biens et services.`

Nous retrouvons les mêmes tendances que pour la variable `Exportations des biens et services`. Les prototypes des classes sont bien séparés. Les bornes inférieure et supérieure du prototype de la classe 1 sont respectivement 79 et 92. L'importation est donc très importante pour ces pays.

Les prototypes des classes pour la variable Taux d'inflation sont représentés à la figure 5.12 :

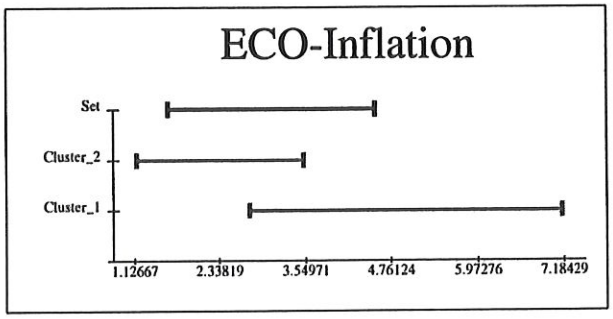

FIG. 5.12 – *Prototypes des classes pour la variable Taux d'inflation.*

Nous voyons ici que les prototypes des classes se recoupent. Cette variable ne permet donc pas de distinguer véritablement les classes. Ceci confirme aussi le fait que le pouvoir discriminant de cette variable est moins important.

Nous pouvons tout de même remarquer que le prototype de la classe 1 est supérieur à celui de la classe 2. Ce qui signifie que, en moyenne, le taux d'inflation est plus élevé dans les pays de la classe 1.

La longueur des intervalles montre qu'il y a une forte évolution du taux d'inflation sur la période d'observation 1999-2003.

Les deux représentations graphiques 5.13 et 5.14 concernent le PIB. La première décrit le PIB par habitant, la seconde son taux d'accroissement. Pour rappel, le pouvoir discriminant de la variable PIB par habitant est nettement plus faible que celui de la variable Taux d'accroissement du PIB.

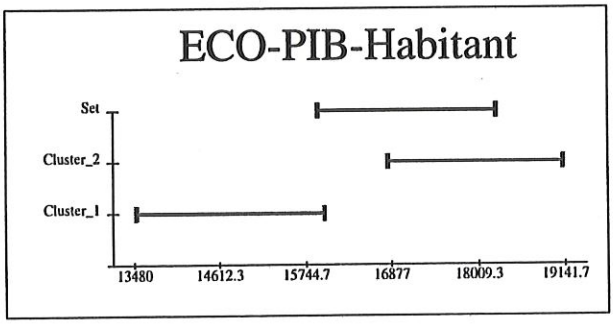

FIG. 5.13 – *Prototypes des classes pour la variable PIB par habitant.*

La représentation graphique des prototypes suivant la valeur du PIB par habitant nous indique que le PIB par habitant des pays de la classe 1 varie en moyenne entre 13.480 et 16.017 USD par habitant.

Il faut être prudent vis-à-vis de cette mesure parce qu'elle camoufle une information importante. Si nous regardons les données de départ, nous constatons que :

- le PIB par habitant se situe entre 40.920 et 45.740 USD pour le Luxembourg ;
- le PIB par habitant se situe entre 21.810 et 27.010 USD pour l'Irlande ;
- le PIB par habitant est inférieur à 10.780 USD pour les autres pays de la classe.

Le Luxembourg et l'Irlande font donc fortement grimper les valeurs moyennes du PIB par habitant pour cette classe. Le prototype de cette classe ne reflète donc pas vraiment la réalité.

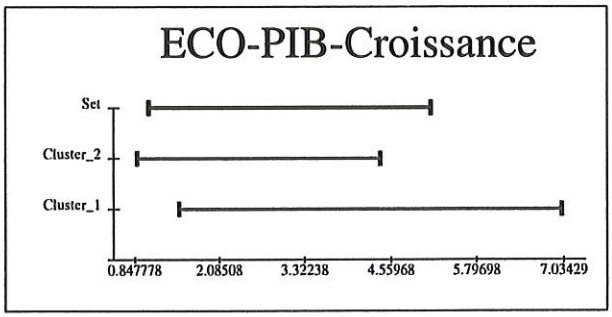

FIG. 5.14 – *Prototypes des classes pour la variable* `Taux de croissance du PIB`.

Dans l'ensemble, nous pouvons dire que les pays de la classe 1 ont tendance à avoir un taux de croissance du PIB plus élevé que ceux de la classe 2. Cette fois encore, nous remarquons que les intervalles sont relativement longs. Ce qui signifie qu'il y a de fortes variations sur la période d'observation 1999-2003 au niveau des taux de croissance du PIB, surtout entre les pays de la classe 1.

5.2.3.3 Conclusion

La partition des pays de l'Union Européenne en deux classes a ici été construite uniquement sur base des indicateurs de développement économiques.

La partition optimale trouvée par SCLUST permet de séparer l'Estonie, la Hongrie, l'Irlande, le Luxembourg, Malte, la République Slovaque et la Slovénie des autres pays de l'Union Européenne. Ces 7 pays se démarquent des autres principalement au niveau de l'importance de l'exportation et de l'importation des biens et services au sein de ces pays. Ces pays ont également un taux de d'inflation assez élevé.

5.2.4 Classification des pays sur base des indicateurs démographiques

Dans cette section, nous nous intéressons uniquement aux indicateurs démographiques.

5.2.4.1 Résultats de la méthode SCLUST

Cette fois encore, nous avons normalisé les données. La partition optimale trouvée par SCLUST est présentée ci-dessous :

```
Classe :   1  Cardinal :   19
==============================

(  0) Allemagne       [0.6]  (  1) Autriche       [0.4]  (  2) Belgique       [0.1]
(  3) Chypre          [1.3]  (  4) Danemark       [0.5]  (  5) Espagne        [1.2]
(  7) Finlande        [0.7]  (  8) France         [0.9]  (  9) Grece          [0.7]
( 11) Irlande         [3.6]  ( 12) Italie         [1.3]  ( 15) Luxembourg     [0.8]
( 16) Malte           [0.8]  ( 17) Pays-Bas       [0.3]  ( 19) Portugal       [0.4]
( 21) Rep.Tcheque     [2.8]  ( 22) Royaume-Uni    [0.4]  ( 23) Slovenie       [1.4]
( 24) Suede           [0.7]

Classe :   2  Cardinal :    6
==============================

(  6) Estonie         [0.4]  ( 10) Hongrie        [2.6]  ( 13) Lettonie       [1.4]
( 14) Lituanie        [0.2]  ( 18) Pologne        [0.8]  ( 20) Rep.Slovaque   [0.5]
```

Cette partition en deux classes sépare l'Union Européenne en deux parties. D'un côté nous retrouvons uniquement six des huit pays de l'Europe de l'Est. De l'autre côté, nous retrouvons tous les pays de l'Europe de l'Est.

Seuls la République Tchèque et la Slovénie se trouvent parmi les pays de l'Europe de l'Ouest, mais comme nous pouvons le constater, ils sont assez éloignés du prototype de leur classe d'appartenance.

La contribution des variables à la formation de la partition est reprise dans le tableau 5.10 :

Liste de variables	Pouvoir discriminant par rapport à la partition	Contribution relative à l'inertie intra-classe
Taux de croissance de la population	35.90	16.42
Taux de fertilité	22.14	10.13
Taux de mortalité infantile	80.85	36.98
Espérance de vie à la naissance	79.73	36.47

TAB. 5.10 – *Contribution de chaque variable.*

En lisant ce tableau, nous remarquons que les variables les plus discriminantes sont les variables `Taux de mortalité infantile` et `Espérance de vie à la naissance`. Le pouvoir discriminant de la variable `Taux de croissance de la population` est moins marqué, de même que celui de la variable `Taux de fertilité`.

Si nous regardons le tableau de données, nous constatons que :

- le taux de mortalité infantile des pays de l'Est est nettement supérieur à celui des pays de l'Ouest ;
- l'espérance de vie à la naissance dans les pays de l'Est est inférieur à celle dans les pays de l'Ouest ;
- le taux d'accroissement de la population est négatif voire nul pour les pays de l'Est, alors qu'il est positif pour les pays de l'Ouest ;
- le taux de fertilité est légèrement inférieur pour les pays de l'Europe de l'Est.

5.2.4.2 Description des classes

Les prototypes des classes suivant les variables `Taux de mortalité infantile` et `Espérance de vie à la naissance` sont représentés aux figures 5.15 et 5.16.

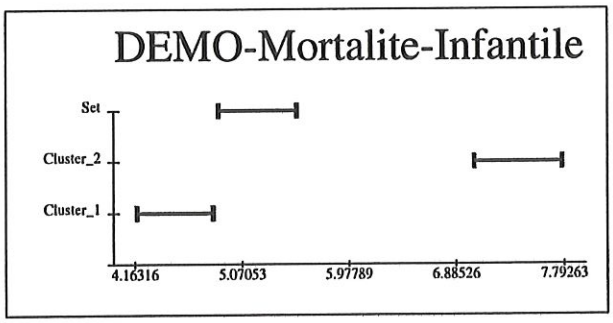

FIG. 5.15 – *Prototypes des classes pour la variable `Taux de mortalité infantile`.*

La variable `Taux de mortalité infantile` permet de séparer les pays de l'Union Européenne en deux classes. Il y a une nette différence entre les taux de mortalité infantile des deux classes. Les pays de l'Europe de l'Est de la classe 2 ont un taux de mortalité infantile qui varie en moyenne entre 8 et 9 enfants sur 1000. Les pays de la classe 1 ont un taux de mortalité infantile qui varie en moyenne entre 4 et 5 enfants sur 1000.

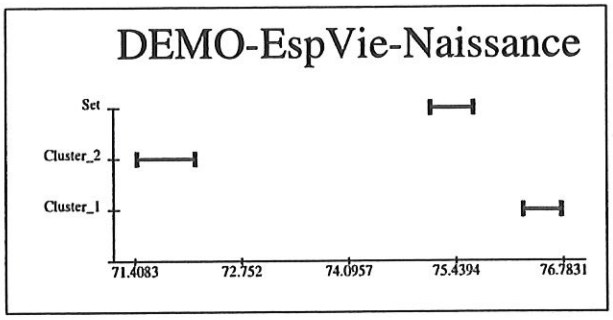

FIG. 5.16 – *Prototypes des classes pour la variable Espérance de vie à la naissance.*

Ici aussi, nous constatons une réelle différence entre les prototypes des classes. L'espérance de vie à la naissance est estimée en moyenne entre 71 et 72 ans pour les pays de la classe 2, entre 77 et 78 ans pour les pays de la classe 1.

Les prototypes des classes suivant la variable `Taux de croissance de la population` sont représentés à la figure 5.17 :

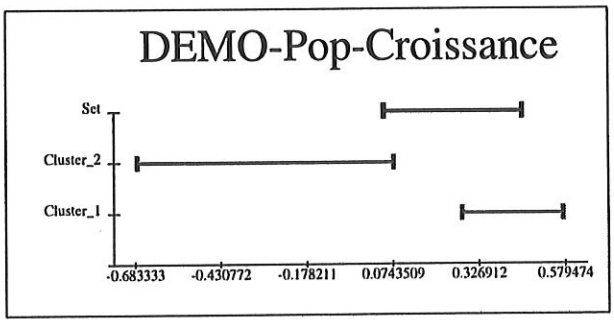

FIG. 5.17 – *Prototypes des classes pour la variable Taux de croissance de la population.*

Pour les pays de la classe 2, le taux de croissance moyen varie entre -0.68 et 0.08. Si nous consultons le tableau de données initiales, nous constatons que les taux de croissance de la population pour les pays de la classe 2 sont négatifs, excepté pour la Hongrie et pour la République Slovaque. Pour les pays de la classe 1, nous constatons qu'il y a un léger accroissement de la population (entre 0,27% et 0,58%).

Les prototypes des classes suivant la variable `Taux de fertilité` sont représentés à la figure 5.18 :

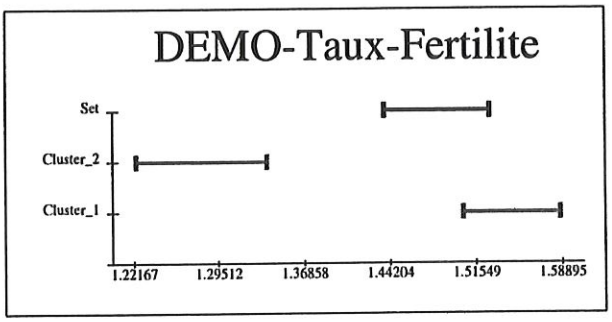

FIG. 5.18 – *Prototypes des classes pour la variable `Taux de fertilité`.*

Il faut être prudent lors de l'interprétation de ce graphique. Les prototypes des classes sont assez bien séparés suivant la variable mais l'échelle de l'axe horizontal est petit. Si nous consultons les données de départ, nous constatons que, dans l'ensemble, les taux de fertilité des pays de la classe 2 sont légèrement inférieurs à ceux de la classe 1.

5.2.4.3 Conclusion

Si nous considérons uniquement les indicateurs démographiques, nous constatons que SCLUST isole les pays baltes, la Hongrie, la Pologne et la République Slovaque des autres pays de l'Union Européenne.

Ces pays se distinguent des autres pays européens surtout au niveau du taux de mortalité infantile. Dans ces pays, en moyenne, entre 8 et 9 enfants sur 1000 décèdent avant l'âge d'un an.

Lors de notre étude, nous avons remarqué que l'espérance de vie à la naissance est en moyenne plus faible dans les pays de l'Est. Nous avons également constaté que l'accroissement de la population sur la période d'observation 1999-2003 est négatif pour la plupart des pays de l'Est, tandis que la population est en légère hausse pour les pays de l'Europe de l'Ouest.

5.2.5 Conclusion de l'étude

Dans le cadre de cette étude, nous nous sommes intéressés aux vingt-cinq pays membres de l'Union Européenne. L'objectif de notre étude était de déterminer si aujourd'hui encore il est possible de percevoir une division de l'Europe en deux parties: Europe de l'Est et Europe de l'Ouest.

En considérant toutes les variables, nous avons constaté que la majorité des pays de l'Europe de l'Ouest (Allemagne, Autriche, Belgique, Finlande, France, Irlande, Luxembourg, Pays-Bas, Royaume-Uni et Suède) étaient toujours classés ensemble. Ceci nous conduit à la conclusion qu'il y a des ressemblances assez fortes entre ces pays au niveau du développement.

Au niveau économique, nous avons constaté que les pays de l'Europe de l'Est se démarquaient des pays de l'Europe de l'Ouest surtout au niveau de l'exportation et de l'importation des biens et services. Nous avons également remarqué une nette différence entre l'Est et l'Ouest de l'Europe au niveau du Produit Intérieur Brut. La valeur en USD du PIB par habitant est extrêmement faible dans les pays de l'Est.

Au niveau démographique, nous avons observé que le taux de mortalité infantile était plus élevé pour les pays de l'Est. Près de 8 enfants sur 1000 meurent avant l'âge d'un an. L'espérance de vie à la naissance est en moyenne inférieure dans les pays de l'Europe de l'Est. Le taux d'accroissement de la population est en légère hausse pour les pays de l'Ouest, contrairement à certains pays de l'Est, pour qui le taux d'accroissement de la population est négatif.

Finalement, nous pouvons dire qu'aujourd'hui encore, des différences assez importantes existent entre les pays de l'Est et de l'Ouest de l'Europe.

Conclusion générale

Le but de ce mémoire fut de réaliser une étude comparative de quatre méthodes de classification symboliques incluses dans le logiciel SODAS 2 développé dans le cadre du projet européen ASSO.

Nous nous sommes tout d'abord intéressés à des ensembles de données symboliques générés aléatoirement de manière à mettre en évidence des structures variées. Dans l'ensemble, les méthodes de classification symboliques ont retrouvé la partition naturelle des données, excepté la méthode SCLASS pour laquelle les résultats différaient parfois.

Nous avons ensuite exposé deux applications. La première traitait d'un problème rencontré en logopédie, lequel nous a été posé par le Docteur Jamart de la Clinique Universitaire de Mont-Godinne. Pour analyser ce problème, nous avons utilisé une approche symbolique et appliqué la méthode de partitionnement SCLUST. Les résultats obtenus étaient tout à fait satisfaisants dans la mesure où ils complétaient l'analyse classique réalisée auparavant. D'ailleurs, des analyses de données symboliques de ce genre pourraient faire progresser la recherche dans le domaine médical.

La seconde application concernait les vingt-cinq pays membres de l'Union Européenne. L'objectif était de déterminer si des différences d'ordre économique, démographique ou autres entre les pays de l'Europe de l'Est et ceux de l'Europe de l'Ouest sont encore perceptibles de nos jours. Appliquées à cet exemple, les quatre méthodes de classification symboliques étudiées donnaient des résultats forts semblables. En effet, lors de l'analyse des partitions en deux classes fournies par les quatre méthodes sur base de toutes les variables, nous avons constaté que dix pays de l'Europe de l'Ouest étaient toujours regroupés au sein d'une même classe, alors que la grande majorité des composants de l'autre classe étaient des pays de l'Europe de l'Est. En prenant en compte des variables purement économiques, la distinction était moins prononcée. Par contre, la considération des variables démographiques apportait une séparation plus nette entre l'Est et l'Ouest.

Annexes

- Annexe 1 : Questionnaire du *Voice Handicap Index*
- Annexe 2 : VHI - Tableau de données classiques
- Annexe 3 : VHI - Tableau de données symboliques
- Annexe 4 : Union Européenne - Tableau de données symboliques

Annexe 1 : Questionnaire du *Voice Handicap Index*

1. (F) On m'entend difficilement à cause de ma voix.
2. (P) Je suis à court de souffle quand je parle.
3. (F) On me comprend difficilement dans un milieu bruyant.
4. (P) Le son de ma voix varie en cours de journée.
5. (F) Les membres de la famille ont du mal à entendre quand je les appelle dans la maison.
6. (F) Je téléphone moins souvent que je le voudrais.
7. (E) Je suis tendu(e) quand je parle avec d'autres à cause de ma voix.
8. (F) J'ai tendance à éviter les groupes de gens à cause de ma voix.
9. (E) Les gens semblent irrités par ma voix.
10. (P) On me demande : "Qu'est-ce qui ne va pas avec ta voix?"
11. (F) Je parle moins souvent avec mes amis, mes voisins, ma famille à cause de ma voix.
12. (F) On me demande de me répéter quand je dialogue face à face avec quelqu'un.
13. (P) Ma voix semble "grinçante" et sèche.
14. (P) J'ai l'impression que je dois forcer pour produire la voix.
15. (E) Je trouve que les autres personnes ne comprennent pas mon problème de voix.
16. (F) Mes difficultés de voix limitent ma vie personnelle et sociale.
17. (P) La clarté de ma voix est imprévisible.
18. (P) J'essaye de changer ma voix pour qu'elle sonne différemment.
19. (F) Je me sens écarté(e) des conversations à cause de ma voix.
20. (P) Je fais beaucoup d'effort pour parler.
21. (P) Ma voix est plus mauvaise le soir.
22. (F) Mes problèmes de voix entraînent des pertes de revenus.
23. (E) Mon problème de voix me contrarie.
24. (E) Je suis moins sociable à cause de mon problème de voix.
25. (E) Je me sens handicapé(e) à cause de ma voix.
26. (P) Ma voix "m'abandonne" au milieu de la conversation.
27. (E) Je suis agacé(e) quand les gens me demandent de répéter.
28. (E) Je suis embarrassé(e) quand les gens me demandent de me répéter.
29. (E) A cause de ma voix, je me sens incompétent.
30. (E) Je suis honteux(se) de mon problème de voix.

Annexe 2 : VHI - Tableau de données classiques

NUM	D1	D2	AGE	SEXE	Z1	Z2	Z3	F1	F2	E1	E2	P1	P2	VHI1	VHI2
1	19/12/2003	26/12/2003	58	0	2	1	1	23	21	23	20	26	23	74	65
2	19/01/2004	26/01/2004	38	0	2	1	1	13	16	21	20	18	23	54	60
3	11/03/2004	30/03/2004	35	0	2	1	1	28	28	26	23	26	23	80	76
4	10/02/2004	10/03/2004	35	0	2	1	1	24	21	21	18	23	25	69	65
5	20/04/2004	27/04/2004	47	0	2	1	1	15	13	15	13	23	22	55	50
6	20/04/2004	28/04/2004	27	1	2	1	1	12	18	17	23	14	11	45	52
7	3/03/2004	10/03/2004	42	0	2	1	1	13	16	19	19	24	22	57	58
8	19/05/2004	24/05/2004	17	0	2	1	1	8	9	6	6	9	8	23	23
9	28/01/2004	15/02/2004	38	1	2	1	1	16	12	13	16	15	14	45	43
10	13/02/2004	8/03/2004	18	0	2	1	1	14	12	17	14	15	13	48	41
11	17/05/2004	7/06/2004	27	0	2	1	1	28	30	26	23	28	31	83	84
12	28/05/2003	25/06/2003	28	0	2	1	1	15	20	19	16	18	20	53	58
13	16/06/2003	24/06/2003	26	0	2	1	1	20	25	14	20	29	28	63	74
14	22/06/2003	29/06/2003	73	0	2	1	1	20	22	19	23	23	20	64	67
15	13/11/2003	18/11/2003	27	0	2	2	1	6	6	22	24	13	9	43	40
16	13/02/2004	20/05/2004	44	0	2	2	1	8	8	16	16	26	26	52	52
17	20/02/2004	27/04/2004	33	0	2	2	1	16	10	17	8	16	12	51	32
18	12/05/2003	2/06/2003	41	0	2	2	1	13	8	11	7	15	16	39	31
19	16/06/2003	6/07/2003	48	1	2	2	1	14	11	14	13	10	8	39	33
20	26/06/2003	2/07/2003	31	1	2	2	1	17	18	15	18	18	18	51	54
21	7/02/2004	26/02/2004	60	1	2	1	2	26	18	22	25	25	25	73	69
22	10/02/2004	26/02/2004	40	0	2	1	2	34	30	28	26	26	23	90	81
23	1/03/2004	8/03/2004	52	1	2	1	2	15	16	11	9	19	22	48	49
24	21/02/2004	10/03/2004	39	0	2	1	2	15	13	5	4	19	20	40	40
25	10/03/2004	28/03/2004	65	1	2	1	2	24	22	19	21	27	25	71	69
26	10/05/2004	19/05/2004	47	1	2	1	2	16	16	25	18	23	23	68	61
27	7/02/2004	21/02/2004	62	0	2	1	2	20	20	21	18	26	22	70	62
28	12/05/2003	2/06/2003	72	1	2	1	2	20	16	7	2	7	10	34	28
29	17/05/2003	9/06/2003	46	0	2	1	2	15	17	4	5	11	16	30	39
30	28/05/2003	20/06/2003	14	0	2	1	2	24	26	24	25	31	31	83	86
31	2/06/2003	11/06/2003	72	0	2	1	2	10	11	13	6	15	10	38	27
32	6/06/2003	15/06/2003	45	0	2	1	2	18	20	26	25	19	22	65	69
33	6/06/2003	20/06/2003	64	1	2	1	2	17	16	3	6	17	16	37	39
34	23/06/2003	30/06/2003	48	0	2	1	2	20	18	15	15	24	26	61	61
35	23/06/2003	30/06/2003	38	0	2	1	2	23	26	15	16	25	26	66	70
36	24/06/2003	7/07/2003	61	0	2	1	2	30	22	22	20	31	26	85	70
37	27/06/2003	4/07/2003	47	0	2	1	2	15	21	4	6	21	25	40	54
38	21/01/2004	4/02/2004	51	0	1	1	1	7	10	2	3	6	7	15	20
39	21/01/2004	2/02/2004	49	0	1	1	1	11	14	11	10	16	17	39	43
40	27/01/2004	11/02/2004	17	0	1	1	1	7	5	4	3	4	6	15	14
41	27/01/2004	10/02/2004	22	0	1	1	1	5	5	1	1	11	4	17	10
42	27/01/2004	10/02/2004	22	0	1	1	1	2	3	0	2	10	7	12	12
43	27/01/2004	10/02/2004	26	0	1	1	1	9	9	4	7	14	12	27	28
44	27/01/2004	10/02/2004	28	0	1	1	1	10	6	2	1	10	8	22	15
45	27/01/2004	10/02/2004	56	0	1	1	1	12	10	5	2	7	10	26	22
46	26/01/2004	10/02/2004	30	0	1	1	1	11	11	9	6	11	11	32	29
47	27/01/2004	10/02/2004	42	0	1	1	1	13	16	12	13	11	13	37	44
48	27/01/2004	19/02/2004	44	0	1	1	1	15	14	2	2	19	14	36	30
49	29/01/2004	19/02/2004	43	0	1	1	1	10	6	2	1	8	0	20	7
50	29/01/2004	19/02/2004	44	1	1	1	1	15	8	5	1	12	8	34	18
51	29/01/2004	19/02/2004	29	0	1	1	1	7	8	7	7	10	9	25	25
52	27/01/2004	10/02/2004	53	0	1	1	1	6	7	1	1	5	3	13	11
53	7/02/2004	22/02/2004	24	0	1	1	1	8	4	3	1	9	7	20	12
54	26/01/2004	8/02/2004	62	1	1	1	1	18	17	23	23	20	20	62	61
55	17/01/2004	5/02/2004	40	0	1	1	1	9	8	4	5	11	9	24	22
56	17/01/2004	31/01/2004	48	0	1	1	1	17	17	14	11	17	16	48	44
57	17/01/2004	7/02/2004	50	1	1	1	1	12	10	7	10	8	13	29	35
58	17/01/2004	24/01/2004	21	0	1	1	1	7	7	3	2	9	7	19	16
59	20/01/2004	9/02/2004	56	0	1	1	1	9	7	6	4	9	10	24	22
60	20/01/2004	5/02/2004	47	0	1	1	1	7	5	1	1	2	2	10	8
61	24/01/2004	8/02/2004	44	0	1	1	1	6	7	2	2	8	8	16	17
62	4/02/2004	18/02/2004	39	0	1	1	1	8	7	4	1	11	8	24	16
63	6/02/2004	18/02/2004	41	0	1	1	1	15	16	8	6	19	15	44	39
64	21/01/2002	4/02/2004	40	0	1	1	1	3	3	0	0	4	3	8	8
65	23/01/2002	9/02/2004	19	0	1	1	1	9	5	0	0	9	4	18	9
66	9/02/2004	16/02/2004	37	0	1	1	1	16	16	4	2	9	10	29	28
67	23/01/2004	16/02/2004	47	0	1	1	1	14	14	2	1	10	13	27	29
68	23/01/2004	14/02/2004	48	1	1	1	1	10	11	10	14	10	11	30	36
69	28/01/2004	11/02/2004	19	1	1	1	1	9	8	5	1	14	13	29	22
70	2/02/2004	11/02/2004	46	1	1	1	1	14	14	17	17	15	19	48	52
71	14/02/2004	1/03/2004	54	0	1	1	1	6	6	1	0	12	6	19	12
72	19/01/2004	2/02/2004	30	0	1	1	1	15	17	11	17	18	19	44	55
73	19/01/2004	2/02/2004	49	0	1	1	1	22	18	15	15	25	21	63	56
74	19/01/2004	2/02/2004	32	0	1	1	1	12	11	2	9	13	13	28	34
75	19/01/2004	2/02/2004	28	1	1	1	1	14	11	3	1	13	8	30	20
76	19/01/2004	2/02/2004	32	0	1	1	1	20	17	2	2	12	11	34	30

NUM	D1	D2	AGE	SEXE	Z1	Z2	Z3	F1	F2	E1	E2	P1	P2	VHI1	VHI2
77	31/01/2004	21/02/2004	19	0	1	1	1	10	10	7	7	7	8	25	26
78	31/01/2004	21/02/2004	18	0	1	1	1	3	5	1	1	1	3	5	9
79	19/01/2004	19/02/2004	42	1	1	1	1	10	9	4	11	8	11	22	32
80	2/02/2004	18/02/2004	35	0	1	1	1	10	11	2	2	17	16	30	30
81	19/01/2004	2/02/2004	22	0	1	1	1	16	17	14	15	11	12	43	45
82	19/01/2004	2/02/2004	60	1	1	1	1	4	4	7	6	5	4	16	14
83	1/02/2004	17/02/2004	44	0	1	1	1	12	11	13	13	5	5	30	29
84	24/01/2004	7/02/2004	46	0	1	1	1	11	10	2	3	16	14	29	27
85	2/02/2004	20/02/2004	28	0	1	1	1	13	15	4	6	16	14	33	35
86	14/01/2004	11/02/2004	36	0	1	1	1	15	18	13	14	14	15	43	48
87	20/01/2004	3/02/2004	54	1	1	1	1	3	1	3	0	8	3	14	4
88	2/02/2004	16/02/2004	32	1	1	1	1	12	11	3	2	12	10	29	24
89	27/01/2004	19/02/2004	47	0	1	1	1	8	8	13	15	9	12	30	35
90	1/02/2004	21/05/2004	46	0	1	1	1	7	11	13	21	8	10	28	42
91	21/01/2004	7/02/2004	46	0	1	2	1	9	10	5	4	10	10	24	24
92	20/01/2004	3/06/2004	32	0	1	2	1	12	8	2	5	12	7	26	20
93	17/01/2004	2/02/2004	50	0	1	2	1	7	7	2	2	3	3	12	12
94	13/01/2004	27/01/2004	28	0	1	2	1	6	8	6	1	6	6	18	15
95	14/01/2004	2/02/2004	67	1	1	2	1	10	10	4	4	2	1	16	15
96	16/01/2004	3/02/2004	36	0	1	2	1	8	4	2	0	3	1	13	5
97	15/01/2004	3/02/2004	45	0	1	2	1	7	9	10	9	15	18	32	36
98	14/01/2004	1/02/2004	43	0	1	2	1	2	3	0	0	5	9	7	12
99	20/01/2004	3/02/2004	26	0	1	2	1	6	6	4	2	4	4	14	12
100	16/01/2004	31/01/2004	29	0	1	2	1	4	7	6	6	4	5	14	18
101	20/01/2004	4/02/2004	35	0	1	2	1	4	7	3	4	3	4	10	15
102	20/01/2004	2/02/2004	62	0	1	2	1	2	5	0	0	8	8	10	13
103	17/01/2004	14/02/2004	57	0	1	2	1	2	5	9	5	6	5	17	15
104	20/01/2004	3/02/2004	36	0	1	2	1	18	13	18	15	13	12	50	41
105	26/01/2004	2/02/2004	40	0	1	2	1	5	5	0	0	1	1	6	6
106	2/02/2004	13/02/2004	34	0	1	2	1	15	15	14	12	20	17	50	45
107	19/01/2004	2/02/2004	40	0	1	2	1	6	5	6	5	4	5	16	15
108	1/04/2004	10/04/2004	58	0	1	2	1	4	5	6	3	4	4	14	12
109	24/05/2004	10/06/2004	43	0	1	2	1	10	9	6	6	9	9	25	24
110	7/02/2004	21/02/2004	55	1	1	1	2	2	5	0	0	0	0	2	5
111	7/02/2004	21/02/2004	24	0	1	1	2	13	13	2	3	8	11	24	27
112	7/02/2004	21/02/2004	59	0	1	1	2	10	10	4	2	7	10	23	24
113	7/02/2004	21/02/2004	64	0	1	1	2	9	7	5	0	16	9	32	16
114	7/02/2004	21/02/2004	45	0	1	1	2	7	7	3	2	9	10	19	19
115	7/02/2004	22/02/2004	57	1	1	1	2	12	7	3	0	10	8	26	17
116	7/02/2004	21/02/2004	64	0	1	1	2	0	4	0	0	2	0	2	4
117	7/02/2004	21/02/2004	44	1	1	1	2	16	12	2	4	15	9	34	26
118	7/02/2004	21/02/2004	55	1	1	1	2	3	3	5	6	8	12	16	21
119	7/02/2004	21/02/2004	48	0	1	1	2	10	9	7	8	14	12	32	29
120	7/02/2004	21/02/2004	53	0	1	1	2	13	8	5	3	17	8	35	19
121	7/02/2004	21/02/2004	53	0	1	1	2	11	12	0	0	11	10	24	22
122	7/02/2004	21/02/2004	48	1	1	1	2	9	6	5	0	10	6	24	12
123	8/02/2004	21/02/2004	55	0	1	1	2	14	15	3	0	15	11	32	26
124	21/02/2004	6/03/2004	34	0	1	1	2	8	12	3	5	15	18	26	35
125	21/02/2004	6/03/2004	27	0	1	1	2	16	13	2	4	22	16	42	35
126	21/02/2004	6/03/2004	67	1	1	1	2	23	22	6	6	16	16	48	47
127	21/02/2004	6/03/2004	76	0	1	1	2	12	14	6	7	6	10	26	33
128	21/02/2004	28/02/2004	62	1	1	1	2	8	8	2	2	3	3	13	13
129	6/03/2004	30/03/2004	37	0	1	1	2	7	6	2	2	15	10	24	18
130	6/03/2004	20/03/2004	47	0	1	1	2	8	6	1	0	5	4	14	10
131	9/02/2004	23/02/2004	57	1	1	1	2	12	6	0	0	3	3	15	9
132	9/02/2004	23/02/2004	52	0	1	1	2	10	10	0	0	16	9	26	19
133	10/01/2004	26/01/2004	24	1	0	0	0	0	0	0	0	1	0	1	0
134	12/01/2004	26/01/2004	30	1	0	0	0	6	20	2	2	6	12	14	34
135	13/01/2004	27/01/2004	60	0	0	0	0	17	14	2	2	14	8	33	24
136	13/01/2004	26/01/2004	62	1	0	0	0	29	22	14	18	22	20	65	60
137	14/01/2004	28/01/2004	44	0	0	0	0	16	10	2	0	10	3	29	13
138	15/01/2004	29/01/2004	49	1	0	0	0	5	4	0	0	0	0	5	4
139	15/01/2004	27/01/2004	38	0	0	0	0	13	5	2	0	9	7	24	12
140	15/01/2004	29/01/2004	48	0	0	0	0	13	6	3	0	12	4	28	10
141	21/01/2004	7/02/2004	45	1	0	0	0	23	15	16	8	21	14	60	37
142	27/01/2004	10/02/2004	27	0	0	0	0	15	12	6	9	14	13	37	34
143	3/02/2004	17/02/2004	37	1	0	0	0	12	9	12	7	12	1	38	17
144	3/02/2004	26/02/2004	37	1	0	0	0	8	2	8	6	0	0	16	8
145	3/02/2004	9/03/2004	33	0	0	0	0	8	17	4	1	5	11	17	31
146	3/02/2004	17/02/2004	28	0	0	0	0	20	19	17	12	19	23	59	57
147	17/02/2004	1/03/2004	22	0	0	0	0	21	17	3	4	15	24	43	47
148	24/01/2004	20/02/2004	22	0	0	0	0	5	3	0	0	1	1	6	4
149	24/05/2004	9/06/2004	23	1	0	0	0	15	12	4	7	14	11	35	32
150	21/02/2004	13/03/2004	46	0	0	0	0	11	13	9	14	19	19	41	46
151	24/05/2004	9/06/2004	23	0	0	0	0	18	16	11	9	18	17	49	44
152	24/05/2004	9/06/2004	57	0	0	0	0	14	16	7	11	14	16	36	44

Annexe 3 : VHI - Tableau de données symboliques

NUM	AGE	SEXE	Z1	Z2	Z3	F	E	P	VHI
1	58	0	2	1	1	[21 ; 23]	[20 ; 23]	[23 ; 26]	[65 ; 74]
2	38	0	2	1	1	[13 ; 16]	[20 ; 21]	[18 ; 23]	[54 ; 60]
3	35	0	2	1	1	[28 ; 28]	[23 ; 26]	[23 ; 26]	[76 ; 80]
4	35	0	2	1	1	[21 ; 24]	[18 ; 21]	[23 ; 25]	[65 ; 69]
5	47	0	2	1	1	[13 ; 15]	[13 ; 15]	[22 ; 23]	[50 ; 55]
6	27	1	2	1	1	[12 ; 18]	[17 ; 23]	[11 ; 14]	[45 ; 52]
7	42	0	2	1	1	[13 ; 16]	[19 ; 19]	[22 ; 24]	[57 ; 58]
8	17	0	2	1	1	[8 ; 9]	[6 ; 6]	[8 ; 9]	[23 ; 23]
9	38	1	2	1	1	[12 ; 16]	[13 ; 16]	[14 ; 15]	[43 ; 45]
10	18	0	2	1	1	[12 ; 14]	[14 ; 17]	[13 ; 15]	[41 ; 48]
11	27	0	2	1	1	[28 ; 30]	[23 ; 26]	[28 ; 31]	[83 ; 84]
12	28	0	2	1	1	[15 ; 20]	[16 ; 19]	[18 ; 20]	[53 ; 58]
13	26	0	2	1	1	[20 ; 25]	[14 ; 20]	[28 ; 29]	[63 ; 74]
14	73	0	2	1	1	[20 ; 22]	[19 ; 23]	[20 ; 23]	[64 ; 67]
15	27	0	2	2	1	[6 ; 6]	[22 ; 24]	[9 ; 13]	[40 ; 43]
16	44	0	2	2	1	[8 ; 8]	[16 ; 16]	[26 ; 26]	[52 ; 52]
17	33	0	2	2	1	[10 ; 16]	[8 ; 17]	[12 ; 16]	[32 ; 51]
18	41	0	2	2	1	[8 ; 13]	[7 ; 11]	[15 ; 16]	[31 ; 39]
19	48	1	2	2	1	[11 ; 14]	[13 ; 14]	[8 ; 10]	[33 ; 39]
20	31	1	2	2	1	[17 ; 18]	[15 ; 18]	[18 ; 18]	[51 ; 54]
21	60	1	2	1	2	[18 ; 26]	[22 ; 25]	[25 ; 25]	[69 ; 73]
22	40	0	2	1	2	[30 ; 34]	[26 ; 28]	[23 ; 26]	[81 ; 90]
23	52	1	2	1	2	[15 ; 16]	[9 ; 11]	[19 ; 22]	[48 ; 49]
24	39	0	2	1	2	[13 ; 15]	[4 ; 5]	[19 ; 20]	[40 ; 40]
25	65	1	2	1	2	[22 ; 24]	[19 ; 21]	[25 ; 27]	[69 ; 71]
26	47	1	2	1	2	[16 ; 16]	[18 ; 25]	[23 ; 23]	[61 ; 68]
27	62	0	2	1	2	[20 ; 20]	[18 ; 21]	[22 ; 26]	[62 ; 70]
28	72	1	2	1	2	[16 ; 20]	[2 ; 7]	[7 ; 10]	[28 ; 34]
29	46	0	2	1	2	[15 ; 17]	[4 ; 5]	[11 ; 16]	[30 ; 39]
30	14	0	2	1	2	[24 ; 26]	[24 ; 25]	[31 ; 31]	[83 ; 86]
31	72	0	2	1	2	[10 ; 11]	[6 ; 13]	[10 ; 15]	[27 ; 38]
32	45	0	2	1	2	[18 ; 20]	[25 ; 26]	[19 ; 22]	[65 ; 69]
33	64	1	2	1	2	[16 ; 17]	[3 ; 6]	[16 ; 17]	[37 ; 39]
34	48	0	2	1	2	[18 ; 20]	[15 ; 15]	[24 ; 26]	[61 ; 61]
35	38	0	2	1	2	[23 ; 26]	[15 ; 16]	[25 ; 26]	[66 ; 70]
36	61	0	2	1	2	[22 ; 30]	[20 ; 22]	[26 ; 31]	[70 ; 85]
37	47	0	2	1	2	[15 ; 21]	[4 ; 6]	[21 ; 25]	[40 ; 54]
38	51	0	1	1	1	[7 ; 10]	[2 ; 3]	[6 ; 7]	[15 ; 20]
39	49	0	1	1	1	[11 ; 14]	[10 ; 11]	[16 ; 17]	[39 ; 43]
40	17	0	1	1	1	[5 ; 7]	[3 ; 4]	[4 ; 6]	[14 ; 15]
41	22	0	1	1	1	[5 ; 5]	[1 ; 1]	[4 ; 11]	[10 ; 17]
42	22	0	1	1	1	[2 ; 3]	[0 ; 2]	[7 ; 10]	[12 ; 12]
43	26	0	1	1	1	[9 ; 9]	[4 ; 7]	[12 ; 14]	[27 ; 28]
44	28	0	1	1	1	[6 ; 10]	[1 ; 2]	[8 ; 10]	[15 ; 22]
45	56	0	1	1	1	[10 ; 12]	[2 ; 5]	[7 ; 10]	[22 ; 26]
46	30	0	1	1	1	[11 ; 11]	[6 ; 9]	[11 ; 11]	[29 ; 32]
47	42	0	1	1	1	[13 ; 16]	[12 ; 13]	[11 ; 13]	[37 ; 44]
48	44	0	1	1	1	[14 ; 15]	[2 ; 2]	[14 ; 19]	[30 ; 36]
49	43	0	1	1	1	[6 ; 10]	[1 ; 2]	[0 ; 8]	[7 ; 20]
50	44	1	1	1	1	[8 ; 15]	[1 ; 5]	[8 ; 12]	[18 ; 34]
51	29	0	1	1	1	[7 ; 8]	[7 ; 7]	[9 ; 10]	[25 ; 25]
52	53	0	1	1	1	[6 ; 7]	[1 ; 1]	[3 ; 5]	[11 ; 13]
53	24	0	1	1	1	[4 ; 8]	[1 ; 3]	[7 ; 9]	[12 ; 20]
54	62	1	1	1	1	[17 ; 18]	[23 ; 23]	[20 ; 20]	[61 ; 62]
55	40	0	1	1	1	[8 ; 9]	[4 ; 5]	[9 ; 11]	[22 ; 24]
56	48	0	1	1	1	[17 ; 17]	[11 ; 14]	[16 ; 17]	[44 ; 48]
57	50	1	1	1	1	[10 ; 12]	[7 ; 10]	[8 ; 13]	[29 ; 35]
58	21	0	1	1	1	[7 ; 7]	[2 ; 3]	[7 ; 9]	[16 ; 19]
59	56	0	1	1	1	[7 ; 9]	[4 ; 6]	[9 ; 10]	[22 ; 24]
60	47	0	1	1	1	[5 ; 7]	[1 ; 1]	[2 ; 2]	[8 ; 10]
61	44	0	1	1	1	[6 ; 7]	[2 ; 2]	[8 ; 8]	[16 ; 17]
62	39	0	1	1	1	[7 ; 8]	[1 ; 4]	[8 ; 11]	[16 ; 24]
63	41	0	1	1	1	[15 ; 16]	[6 ; 8]	[15 ; 19]	[39 ; 44]
64	40	0	1	1	1	[3 ; 3]	[0 ; 0]	[3 ; 4]	[8 ; 8]
65	19	0	1	1	1	[5 ; 9]	[0 ; 0]	[4 ; 9]	[9 ; 18]
66	37	0	1	1	1	[16 ; 16]	[2 ; 4]	[9 ; 10]	[28 ; 29]
67	47	0	1	1	1	[14 ; 14]	[1 ; 2]	[10 ; 13]	[27 ; 29]
68	48	1	1	1	1	[10 ; 11]	[10 ; 14]	[10 ; 11]	[30 ; 36]
69	19	1	1	1	1	[8 ; 9]	[1 ; 5]	[13 ; 14]	[22 ; 29]
70	46	1	1	1	1	[14 ; 14]	[17 ; 17]	[15 ; 19]	[48 ; 52]
71	54	0	1	1	1	[6 ; 6]	[0 ; 1]	[6 ; 12]	[12 ; 19]
72	30	0	1	1	1	[15 ; 17]	[11 ; 17]	[18 ; 19]	[44 ; 55]
73	49	0	1	1	1	[18 ; 22]	[15 ; 15]	[21 ; 25]	[56 ; 63]
74	32	0	1	1	1	[11 ; 12]	[2 ; 9]	[13 ; 13]	[28 ; 34]
75	28	1	1	1	1	[11 ; 14]	[1 ; 3]	[8 ; 13]	[20 ; 30]
76	32	0	1	1	1	[17 ; 20]	[2 ; 2]	[11 ; 12]	[30 ; 34]

NUM	AGE	SEXE	Z1	Z2	Z3	F	E	P	VHI
77	19	0	1	1	1	[10 ; 10]	[7 ; 7]	[7 ; 8]	[25 ; 26]
78	18	0	1	1	1	[3 ; 5]	[1 ; 1]	[1 ; 3]	[5 ; 9]
79	42	1	1	1	1	[9 ; 10]	[4 ; 11]	[8 ; 11]	[22 ; 32]
80	35	0	1	1	1	[10 ; 11]	[2 ; 2]	[16 ; 17]	[30 ; 30]
81	22	0	1	1	1	[16 ; 17]	[14 ; 15]	[11 ; 12]	[43 ; 45]
82	60	1	1	1	1	[4 ; 4]	[6 ; 7]	[4 ; 5]	[14 ; 16]
83	44	0	1	1	1	[11 ; 12]	[13 ; 13]	[5 ; 5]	[29 ; 30]
84	46	0	1	1	1	[10 ; 11]	[2 ; 3]	[14 ; 16]	[27 ; 29]
85	28	0	1	1	1	[13 ; 15]	[4 ; 6]	[14 ; 16]	[33 ; 35]
86	36	0	1	1	1	[15 ; 18]	[13 ; 14]	[14 ; 15]	[43 ; 48]
87	54	1	1	1	1	[1 ; 3]	[0 ; 3]	[3 ; 8]	[4 ; 14]
88	32	1	1	1	1	[11 ; 12]	[2 ; 3]	[10 ; 12]	[24 ; 29]
89	47	0	1	1	1	[8 ; 8]	[13 ; 15]	[9 ; 12]	[30 ; 35]
90	46	0	1	1	1	[7 ; 11]	[13 ; 21]	[8 ; 10]	[28 ; 42]
91	46	0	1	2	1	[9 ; 10]	[4 ; 5]	[10 ; 10]	[24 ; 24]
92	32	0	1	2	1	[8 ; 12]	[2 ; 5]	[7 ; 12]	[20 ; 26]
93	50	0	1	2	1	[7 ; 7]	[2 ; 2]	[3 ; 3]	[12 ; 12]
94	28	0	1	2	1	[6 ; 8]	[1 ; 6]	[6 ; 6]	[15 ; 18]
95	67	1	1	2	1	[10 ; 10]	[4 ; 4]	[1 ; 2]	[15 ; 16]
96	36	0	1	2	1	[4 ; 8]	[0 ; 2]	[1 ; 3]	[5 ; 13]
97	45	0	1	2	1	[7 ; 9]	[9 ; 10]	[15 ; 18]	[32 ; 36]
98	43	0	1	2	1	[2 ; 3]	[0 ; 0]	[5 ; 9]	[7 ; 12]
99	26	0	1	2	1	[6 ; 6]	[2 ; 4]	[4 ; 4]	[12 ; 14]
100	29	0	1	2	1	[4 ; 7]	[6 ; 6]	[4 ; 5]	[14 ; 18]
101	35	0	1	2	1	[4 ; 7]	[3 ; 4]	[3 ; 4]	[10 ; 15]
102	62	0	1	2	1	[2 ; 5]	[0 ; 0]	[8 ; 8]	[10 ; 13]
103	57	0	1	2	1	[2 ; 5]	[5 ; 9]	[5 ; 6]	[15 ; 17]
104	36	0	1	2	1	[13 ; 18]	[15 ; 18]	[12 ; 13]	[41 ; 50]
105	40	0	1	2	1	[5 ; 5]	[0 ; 0]	[1 ; 1]	[6 ; 6]
106	34	0	1	2	1	[15 ; 15]	[12 ; 14]	[17 ; 20]	[45 ; 50]
107	40	0	1	2	1	[5 ; 6]	[5 ; 6]	[4 ; 5]	[15 ; 16]
108	58	0	1	2	1	[4 ; 5]	[3 ; 6]	[4 ; 4]	[12 ; 14]
109	43	0	1	2	1	[9 ; 10]	[6 ; 6]	[9 ; 9]	[24 ; 25]
110	55	1	1	1	2	[2 ; 5]	[0 ; 0]	[0 ; 0]	[2 ; 5]
111	24	0	1	1	2	[13 ; 13]	[2 ; 3]	[8 ; 11]	[24 ; 27]
112	59	0	1	1	2	[10 ; 10]	[2 ; 4]	[7 ; 10]	[23 ; 24]
113	64	0	1	1	2	[7 ; 9]	[0 ; 5]	[9 ; 16]	[16 ; 32]
114	45	0	1	1	2	[7 ; 7]	[2 ; 3]	[9 ; 10]	[19 ; 19]
115	57	1	1	1	2	[7 ; 12]	[0 ; 3]	[8 ; 10]	[17 ; 26]
116	64	0	1	1	2	[0 ; 4]	[0 ; 0]	[0 ; 2]	[2 ; 4]
117	44	1	1	1	2	[12 ; 16]	[2 ; 4]	[9 ; 15]	[26 ; 34]
118	55	1	1	1	2	[3 ; 3]	[5 ; 6]	[8 ; 12]	[16 ; 21]
119	48	0	1	1	2	[9 ; 10]	[7 ; 8]	[12 ; 14]	[29 ; 32]
120	53	0	1	1	2	[8 ; 13]	[3 ; 5]	[8 ; 17]	[19 ; 35]
121	53	0	1	1	2	[11 ; 12]	[0 ; 0]	[10 ; 11]	[22 ; 24]
122	48	1	1	1	2	[6 ; 9]	[0 ; 5]	[6 ; 10]	[12 ; 24]
123	55	0	1	1	2	[14 ; 15]	[0 ; 3]	[11 ; 15]	[26 ; 32]
124	34	0	1	1	2	[8 ; 12]	[3 ; 5]	[15 ; 18]	[26 ; 35]
125	27	0	1	1	2	[13 ; 16]	[2 ; 4]	[16 ; 22]	[35 ; 42]
126	67	1	1	1	2	[22 ; 23]	[6 ; 6]	[16 ; 16]	[47 ; 48]
127	76	0	1	1	2	[12 ; 14]	[6 ; 7]	[6 ; 10]	[26 ; 33]
128	62	1	1	1	2	[8 ; 8]	[2 ; 2]	[3 ; 3]	[13 ; 13]
129	37	0	1	1	2	[6 ; 7]	[2 ; 2]	[10 ; 15]	[18 ; 24]
130	47	0	1	1	2	[6 ; 8]	[0 ; 1]	[4 ; 5]	[10 ; 14]
131	57	1	1	1	2	[6 ; 12]	[0 ; 0]	[3 ; 3]	[9 ; 15]
132	52	0	1	1	2	[10 ; 10]	[0 ; 0]	[9 ; 16]	[19 ; 26]
133	24	1	0	0	0	[0 ; 0]	[0 ; 0]	[0 ; 1]	[0 ; 1]
134	30	1	0	0	0	[6 ; 20]	[2 ; 2]	[6 ; 12]	[14 ; 34]
135	60	0	0	0	0	[14 ; 17]	[2 ; 2]	[8 ; 14]	[24 ; 33]
136	62	1	0	0	0	[22 ; 29]	[14 ; 18]	[20 ; 22]	[60 ; 65]
137	44	0	0	0	0	[10 ; 16]	[0 ; 2]	[3 ; 10]	[13 ; 29]
138	49	1	0	0	0	[4 ; 5]	[0 ; 0]	[0 ; 0]	[4 ; 5]
139	38	0	0	0	0	[5 ; 13]	[0 ; 2]	[7 ; 9]	[12 ; 24]
140	48	0	0	0	0	[6 ; 13]	[0 ; 3]	[4 ; 12]	[10 ; 28]
141	45	1	0	0	0	[15 ; 23]	[8 ; 16]	[14 ; 21]	[37 ; 60]
142	27	0	0	0	0	[12 ; 15]	[6 ; 9]	[13 ; 14]	[34 ; 37]
143	37	1	0	0	0	[9 ; 12]	[7 ; 12]	[1 ; 12]	[17 ; 38]
144	37	1	0	0	0	[2 ; 8]	[6 ; 8]	[0 ; 0]	[8 ; 16]
145	33	0	0	0	0	[8 ; 17]	[1 ; 4]	[5 ; 11]	[17 ; 31]
146	28	0	0	0	0	[19 ; 20]	[12 ; 17]	[19 ; 23]	[57 ; 59]
147	22	0	0	0	0	[17 ; 21]	[3 ; 4]	[15 ; 24]	[43 ; 47]
148	22	0	0	0	0	[3 ; 5]	[0 ; 0]	[1 ; 1]	[4 ; 6]
149	23	1	0	0	0	[12 ; 15]	[4 ; 7]	[11 ; 14]	[32 ; 35]
150	46	0	0	0	0	[11 ; 13]	[9 ; 14]	[19 ; 19]	[41 ; 46]
151	23	0	0	0	0	[16 ; 18]	[9 ; 11]	[17 ; 18]	[44 ; 49]
152	57	0	0	0	0	[14 ; 16]	[7 ; 11]	[14 ; 16]	[36 ; 44]

Annexe 4 : Union Européenne - Tableau de données symboliques

	ECONOMIE					DEMOGRAPHIE			
Pays	PIB par hab.	Croiss.PIB	Exportation	Importation	Inflation	Croiss.pop.	Fertilité	Mort.infant.	Esp.vie naiss.
All.	[22860 ; 25730]	[-0,10 ; 2,86]	[29,64 ; 35,95]	[28,83 ; 33,43]	[-0,25 ; 1,59]	[0,04 ; 0,21]	[1,30 ; 1,36]	[4,20 ; 4,50]	[77,73 ; 78,38]
Aut.	[23970 ; 26810]	[0,75 ; 3,42]	[45,48 ; 52,75]	[46,34 ; 52,25]	[0,70 ; 2,06]	[0,19 ; 0,42]	[1,33 ; 1,40]	[4,10 ; 4,80]	[77,93 ; 79,06]
Bel.	[22960 ; 25760]	[0,64 ; 3,84]	[75,71 ; 85,82]	[71,37 ; 82,60]	[1,23 ; 1,78]	[0,23 ; 0,45]	[1,61 ; 1,66]	[4,00 ; 5,00]	[77,38 ; 78,30]
Chy.	[12220 ; 12460]	[2,00 ; 5,10]	[44,51 ; 44,51]	[48,32 ; 48,32]	[2,22 ; 4,50]	[0,40 ; 0,67]	[1,90 ; 1,90]	[4,00 ; 6,00]	[78,09 ; 78,16]
Dan.	[29880 ; 33570]	[0,43 ; 2,83]	[38,10 ; 44,67]	[33,32 ; 38,68]	[1,62 ; 2,98]	[0,24 ; 0,39]	[1,72 ; 1,77]	[4,20 ; 5,30]	[76,48 ; 77,14]
Esp.	[14490 ; 17040]	[2,04 ; 4,22]	[27,51 ; 30,12]	[28,77 ; 32,36]	[2,75 ; 4,44]	[0,45 ; 0,87]	[1,20 ; 1,26]	[4,00 ; 4,00]	[78,68 ; 79,56]
Est.	[3800 ; 5380]	[-0,08 ; 7,80]	[72,20 ; 88,26]	[76,80 ; 92,11]	[2,40 ; 5,77]	[-0,76 ; -0,37]	[1,23 ; 1,37]	[8,00 ; 8,00]	[70,51 ; 71,18]
Fin.	[23970 ; 27060]	[1,07 ; 5,12]	[37,01 ; 42,99]	[29,26 ; 33,74]	[-0,21 ; 3,19]	[0,14 ; 0,31]	[1,72 ; 1,76]	[3,00 ; 3,70]	[77,29 ; 78,32]
Fra.	[22180 ; 24730]	[0,47 ; 3,79]	[25,81 ; 28,55]	[23,68 ; 27,31]	[0,54 ; 2,26]	[0,38 ; 0,50]	[1,79 ; 1,89]	[4,30 ; 4,50]	[78,51 ; 79,31]
Gre.	[11010 ; 13230]	[3,42 ; 4,45]	[19,84 ; 25,59]	[27,67 ; 34,14]	[3,01 ; 3,86]	[0,25 ; 0,44]	[1,25 ; 1,29]	[4,00 ; 5,00]	[77,99 ; 77,99]
Hon.	[4490 ; 6350]	[3,05 ; 5,21]	[65,15 ; 74,88]	[67,63 ; 78,73]	[7,83 ; 9,90]	[-0,46 ; 1,61]	[1,29 ; 1,32]	[7,40 ; 9,20]	[70,68 ; 72,58]
Irl.	[21810 ; 27010]	[3,70 ; 11,28]	[87,60 ; 98,43]	[74,22 ; 84,37]	[3,85 ; 5,43]	[1,34 ; 1,67]	[1,88 ; 1,98]	[5,10 ; 5,90]	[76,44 ; 77,69]
Ita.	[19110 ; 21570]	[0,26 ; 3,03]	[25,38 ; 28,39]	[23,49 ; 27,31]	[1,57 ; 3,07]	[-0,08 ; 0,10]	[1,23 ; 1,29]	[4,30 ; 5,10]	[79,12 ; 79,83]
Let.	[2810 ; 4400]	[3,29 ; 8,01]	[43,91 ; 47,04]	[54,24 ; 56,95]	[-1,06 ; 5,29]	[-0,89 ; -0,55]	[1,16 ; 1,29]	[10,00 ; 10,00]	[69,74 ; 70,69]
Lit.	[2910 ; 4500]	[-1,70 ; 8,96]	[39,10 ; 53,91]	[49,24 ; 59,87]	[-0,61 ; 1,35]	[-0,74 ; -0,37]	[1,24 ; 1,35]	[8,00 ; 9,00]	[71,55 ; 71,97]
Lux.	[40920 ; 45740]	[1,35 ; 9,04]	[136,13 ; 153,39]	[119,68 ; 136,14]	[0,71 ; 4,14]	[0,34 ; 1,38]	[1,63 ; 1,78]	[4,50 ; 5,10]	[77,87 ; 78,32]
Mal.	[9270 ; 10780]	[-1,75 ; 6,30]	[87,36 ; 102,75]	[89,00 ; 113,46]	[0,85 ; 5,89]	[0,50 ; 1,27]	[1,41 ; 1,72]	[5,00 ; 6,00]	[78,39 ; 78,53]
P-B.	[23520 ; 26230]	[-0,90 ; 4,00]	[60,25 ; 67,47]	[55,91 ; 62,25]	[1,55 ; 5,39]	[0,48 ; 0,75]	[1,65 ; 1,75]	[4,50 ; 5,20]	[77,83 ; 78,49]
Pol.	[4270 ; 5280]	[1,02 ; 4,10]	[20,96 ; 27,84]	[26,35 ; 34,38]	[0,65 ; 6,57]	[-1,03 ; -0,02]	[1,24 ; 1,37]	[6,00 ; 8,00]	[73,04 ; 74,60]
Por.	[10550 ; 11800]	[-1,20 ; 3,80]	[29,70 ; 31,54]	[37,62 ; 42,77]	[3,08 ; 5,08]	[0,44 ; 0,73]	[1,42 ; 1,52]	[4,00 ; 6,00]	[75,46 ; 76,18]
RSl.	[3860 ; 4940]	[1,47 ; 4,40]	[61,38 ; 78,04]	[65,73 ; 81,54]	[2,57 ; 8,45]	[-0,22 ; 0,20]	[1,17 ; 1,33]	[7,00 ; 8,00]	[72,93 ; 73,39]
RTc.	[5500 ; 7150]	[1,21 ; 3,89]	[56,46 ; 66,49]	[57,65 ; 69,02]	[1,38 ; 4,91]	[-0,48 ; 0,01]	[1,13 ; 1,18]	[3,90 ; 4,60]	[74,60 ; 75,32]
R-U.	[24140 ; 28320]	[1,64 ; 3,78]	[25,11 ; 28,07]	[28,05 ; 30,12]	[1,41 ; 3,27]	[0,17 ; 0,30]	[1,63 ; 1,71]	[5,20 ; 5,80]	[77,59 ; 77,63]
Slo.	[10210 ; 11920]	[2,52 ; 5,21]	[52,41 ; 59,65]	[56,49 ; 60,11]	[0,90 ; 10,71]	[0,05 ; 0,18]	[1,21 ; 1,26]	[4,00 ; 5,00]	[74,93 ; 76,09]
Sue.	[26140 ; 28910]	[0,92 ; 4,58]	[42,63 ; 46,11]	[36,48 ; 40,29]	[0,70 ; 2,31]	[0,06 ; 0,36]	[1,54 ; 1,71]	[2,80 ; 3,70]	[79,43 ; 80,11]

	EDUCATION	CONSOMMATION			COMMUNICATION		
Pays	Education	Electricité	Carburant	Conso.CO2	Téléphone	Ordinateurs	Internet
All.	[87,31 ; 88,32]	[5689,68 ; 6137,03]	[4162,84 ; 4291,94]	[9,55 ; 9,64]	[872,12 ; 1442,48]	[296,97 ; 484,66]	[208,12 ; 472,55]
Aut.	[88,12 ; 88,51]	[6428,55 ; 6838,46]	[3598,48 ; 3842,50]	[7,59 ; 7,65]	[992,16 ; 1359,53]	[256,82 ; 369,29]	[225,02 ; 462,03]
Bel.	[95,36 ; 95,36]	[7286,13 ; 7596,19]	[5505,37 ; 5785,31]	[9,97 ; 10,20]	[820,52 ; 1281,95]	[219,75 ; 318,15]	[136,73 ; 385,64]
Chy.	[88,06 ; 91,66]	[3671,09 ; 4425,03]	[3054,38 ; 3224,98]	[7,97 ; 8,48]	[860,55 ; 1315,88]	[194,32 ; 269,89]	[131,54 ; 337,11]
Dan.	[89,46 ; 92,89]	[6023,67 ; 6076,40]	[3635,58 ; 3759,35]	[8,35 ; 8,96]	[1179,38 ; 1552,55]	[451,67 ; 576,82]	[306,01 ; 512,82]
Esp.	[90,02 ; 94,01]	[4409,03 ; 5047,62]	[2945,85 ; 3215,22]	[6,87 ; 6,99]	[783,15 ; 1342,93]	[119,40 ; 195,95]	[70,39 ; 239,11]
Est.	[83,31 ; 86,79]	[3462,24 ; 3882,18]	[3302,67 ; 3443,55]	[11,63 ; 11,68]	[625,69 ; 1118,61]	[135,19 ; 440,41]	[138,66 ; 444,12]
Fin.	[94,38 ; 95,11]	[14373,48 ; 15326,22]	[6379,16 ; 6851,70]	[10,33 ; 11,02]	[1185,60 ; 1401,56]	[360,11 ; 441,71]	[322,74 ; 533,82]
Fra.	[92,35 ; 93,30]	[6391,62 ; 6681,62]	[4351,93 ; 4500,21]	[6,15 ; 6,23]	[943,68 ; 1261,91]	[267,47 ; 347,10]	[91,60 ; 365,61]
Gre.	[84,03 ; 84,95]	[3732,06 ; 4231,17]	[2446,29 ; 2637,44]	[7,94 ; 8,21]	[895,46 ; 1356,16]	[60,23 ; 81,68]	[70,58 ; 149,99]
Hon.	[86,87 ; 92,11]	[2874,26 ; 3099,12]	[2495,31 ; 2515,89]	[5,40 ; 5,72]	[533,05 ; 1117,41]	[74,67 ; 108,35]	[59,74 ; 232,24]
Irl.	[81,71 ; 82,39]	[4997,26 ; 5555,22]	[3706,62 ; 3917,22]	[10,83 ; 11,08]	[911,69 ; 1370,98]	[315,11 ; 420,76]	[109,49 ; 316,67]
Ita.	[87,59 ; 90,54]	[4535,15 ; 4901,20]	[2958,42 ; 2993,93]	[7,34 ; 7,42]	[990,50 ; 1501,58]	[156,95 ; 230,68]	[143,00 ; 352,44]
Let.	[86,77 ; 88,69]	[1866,53 ; 2088,11]	[1521,92 ; 1824,64]	[2,53 ; 2,72]	[412,34 ; 811,23]	[81,99 ; 188,00]	[43,04 ; 403,59]
Lit.	[92,39 ; 92,85]	[1768,05 ; 1938,02]	[2049,93 ; 2475,93]	[3,39 ; 3,82]	[401,35 ; 868,93]	[59,48 ; 109,75]	[27,85 ; 201,90]
Lux.	[79,71 ; 81,66]	[12754,63 ; 13050,23]	[8083,33 ; 9111,61]	[18,63 ; 19,37]	[1202,64 ; 1991,30]	[393,11 ; 594,17]	[173,43 ; 376,52]
Mal.	[79,95 ; 81,99]	[3907,22 ; 4173,80]	[1868,35 ; 2246,85]	[7,22 ; 9,29]	[609,50 ; 1245,66]	[181,32 ; 255,05]	[77,71 ; 303,03]
P-B.	[89,87 ; 91,88]	[5993,17 ; 6199,14]	[4651,00 ; 4826,75]	[8,72 ; 9,18]	[1031,17 ; 1388,05]	[359,31 ; 466,63]	[390,82 ; 521,95]
Pol.	[90,83 ; 90,83]	[2388,02 ; 2541,32]	[2316,96 ; 2410,65]	[7,80 ; 8,13]	[364,78 ; 769,65]	[61,95 ; 142,01]	[54,21 ; 232,45]
Por.	[84,53 ; 87,32]	[3550,23 ; 4000,10]	[2463,34 ; 2545,52]	[5,85 ; 5,93]	[890,34 ; 1309,63]	[93,02 ; 134,87]	[150,04 ; 280,58]
RSl.	[86,61 ; 86,61]	[4082,73 ; 4359,92]	[3217,73 ; 3447,85]	[6,57 ; 7,18]	[429,69 ; 924,95]	[109,30 ; 180,36]	[54,16 ; 255,87]
RTc.	[88,34 ; 89,48]	[4679,76 ; 4982,16]	[3718,66 ; 4090,29]	[10,59 ; 11,56]	[559,24 ; 1324,94]	[106,97 ; 177,44]	[68,07 ; 308,01]
R-U.	[94,58 ; 95,56]	[5495,17 ; 5639,20]	[3824,28 ; 3969,62]	[9,15 ; 9,64]	[1028,66 ; 1431,34]	[302,52 ; 405,70]	[210,08 ; 423,10]
Slo.	[90,94 ; 92,68]	[5218,33 ; 5907,22]	[3220,85 ; 3485,96]	[7,26 ; 7,34]	[698,39 ; 1341,38]	[251,40 ; 300,60]	[125,70 ; 375,75]
Sue.	[96,10 ; 98,60]	[14290,88 ; 14917,13]	[5355,62 ; 5755,79]	[5,27 ; 5,29]	[1318,52 ; 1624,53]	[451,39 ; 621,27]	[413,70 ; 573,07]

Bibliographie

- Bock, H.H et Diday, E., *Analysis of symbolic data. Exploratory methods extracting statistical information from complex data.* Springer-Verlag, 2000.

- Brito, P., *Hierarchical and pyramidal clustering. Clustering and visualizing symbolic data by using the module* HIPYR *and* VPYR, 2003.

- Celeux, G., Diday, E., Govaert, G., Lechevallier, Y. et Ralanbondrainy H., *Classification automatique des données.* Dunod-Informatique, Bordas, 1989.

- Chavent, M., *Analyse des données symboliques. Une méthode divisive de classification.* Thèse de doctorat, Université Paris IX-Dauphine, 1997.

- Jacobson, B.H., Johnson, A., Grywalsky, C., Silbergleit, A., Jacobson, G., Benninger, M.S. et Newman, C.W. *The voice handicap index (VHI) : Development and validation.* American Journal of speak-language pathology, Vol 6 p.66-70, 1997.

- Pirçon, J.Y., *La classification et les processus de Poisson pour de nouvelles méthodes monothétiques de partitionnement.* Thèse de doctorat, FUNDP Namur, 2004.

- Rasson, J.P., Pirçon, J.Y., Lallemand, P. et Adans, S., *Unsupervised divisive classification.* Publications du Département de Mathématique, FUNDP Namur, 2004.

- Rasson, J.P. et Pirçon, J.Y., *Les arbres de clustering.* Publications du Département de Mathématique, FUNDP Namur, 2002.

- Rosen, C.A. et Murry, T. *Voice handicap in singers.* Journal of voice, Vol 14 No 3 p. 370-377, 2000.

- Simon, C., *Essai d'adaptation d'une échelle d'auto-évaluation de la voix parlée (le voice handicap index de Jacobson, 1997) à la voix chantée. Validation d'une échelle sur une population de 37 chanteurs classiques dysphoniques, 95 chanteurs classiques normophoniques et 20 sujets non chanteurs sans plainte vocale.* Mémoire, Université Catholique de Louvain-Université Libre de Bruxelles, 2004.

- *Tutorial for* SCLUST *module.* ASSO, 2001.

- SCLASS *user manual, Unsupervised classification tree.* Publications du Département de Mathématique, FUNDP Namur, 2000.

- DIV *help guide.* INRIA, 2000.

- HIPYR *and* VPYR *help guide. Hierarchical and pyramidal clustering.* FEP, 2000.

- `www.worldbank.org`