Titre: Title:	Validation d'une heuristique de celeux pour la détermination des partitions centrales
Auteur: Author:	Soumaya Moussa
Date:	1989
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Moussa, S. (1989). Validation d'une heuristique de celeux pour la détermination des partitions centrales [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/58264/

UNIVERSITÉ DE MONTRÉAL

VALIDATION D'UNE HEURISTIQUE DE CELEUX POUR LA DÉTERMINATION DES PARTITIONS CENTRALES

par

Soumaya Moussa
DÉPARTEMENT DE MATHÉMATIQUES APPLIQUÉES
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU GRADE DE MAÎTRE ÈS SCIENCES APPLIQUÉES (M.Sc.A.)
(MATHÉMATIQUES APPLIQUÉES)
Décembre 1989

ISBN 0-315-58192-1

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE

Ce mémoire intitulé:

VALIDATION D'UNE HEURISTIQUE DE CELEUX POUR LA DÉTERMINATION DES PARTITIONS CENTRALES

présenté par: Soumaya Moussa

en vue de l'obtention du grade de: MAÎTRE ÈS SCIENCES APPLIQUÉES (M.Sc.A.)

a été dûment accepté par le jury d'examen constitué de:

M. Bernard Clément ..... Ph.D., président

M. Marc Bourdeau ..... Ph.D.

Mme Pascale Rousseau ..... Ph.D.

À M., O.

et à mes parents

Sommaire

Gilles Celeux (1984) a proposé une heuristique qui offre une solution sous–optimale au problème du consensus d'une multipartition comprenant m partitions d'un ensemble E. Nous proposons une méthode originale de validation de cette heuristique basée sur des idées expérimentales qu'on trouve en contrôle de la qualité.

Nous obtenons nos multipartitions à partir de l'algorithme des centres mobiles, initialisé avec un nombre constant de noyaux. Nous avons travaillé sur des données simulées en dimension deux uniquement, et dont les géométries sont très simples.

Notre technique de validation nous a permis de calculer les effets de certains facteurs sur la performance de l'heuristique de Celeux. Les facteurs explorés sont: la longueur de la multipartition, la nature de la distribution des données simulées, le nombre et la séparation entre les classes, le nombre et la nature des points amorces. Chacun des facteurs a été fixé à deux modalités.

Deux expériences ont été menées dont les principales conclusions sont:

(1) L'heuristique de Celeux est dans l'incapacité de retrouver une structure

simulée avec un seul groupe. Nos validations posent la question des circonstances qui entraînent que la partition grossière soit une partition centrale. Nous formulons une conjecture à ce sujet, et nous la démontrons dans le cas particulier $m = 2$.

(2) L'initialisation sur des noyaux alignés montre une légère supériorité sur des noyaux randomisés dans des structures avec deux groupes.

(3) L'interaction des facteurs: nombre de groupes et nombre de points amorces, a un effet très important sur la performance de l'heuristique. On constate qu'une légère surestimation du nombre de classes donne de meilleurs résultats.

Abstract

Gilles Celeux (1984) has proposed a heuristic for the consensus problem of a multipartition composed of m partitions of a set E, which gives a suboptimal solution for the central partition. We propose a validation study of this heuristic based on some experimental ideas found in Quality Control.

Our simulation studies are built on very simple structures for the data set, constructed in two dimensions only. The multipartitions are obtained through k-means algorithms with fixed number of seed points.

Our validation technique has put forward the effect of certain factors on the performance of the heuristic. The explored factors were: the length of the multipartition, the nature of the distribution for the simulated data, the number and separation of the clusters, the number and the nature of the seed points. Each factor was given two levels.

The main conclusions of our two experiments are:

(1) The heuristic is incapable of finding a one-group structure. We are led to the question of how to get the one-subset partition of E as the optimal or

central partition of a multipartition. We formulate a conjecture for a sufficient and necessary condition, and prove the conjecture in the case $m = 2$.

(2) Initializing the k-means algorithms on aligned seed points gives better results in the case of two-group structures.

(3) The interaction between the factors: number of groups and number of seed points, has a very important effect on the performance of the heuristic. We note that a slight overestimation of the number of classes gives better results.

Remerciements

Qu'il me soit permis de témoigner ma reconnaissance et toute ma gratitude à mon directeur de recherche Monsieur Marc Bourdeau, pour la confiance qu'il m'a témoignée, son entière disponibilité, son soutien moral et financier. C'est grâce à ses prodigues suggestions que j'ai pu mener à bien ce travail.

Je remercie aussi Monsieur Bernard Clément de m'avoir fait bénéficier de ses précieux conseils et de son expérience dans le domaine de la planification d'expériences, et aussi d'avoir bien voulu présider le jury d'examen de mon mémoire.

Je remercie de même Madame Pascale Rousseau, professeur à l'UQAM, d'avoir accepté d'examiner ce travail.

Je suis redevable enfin au département de mathématiques appliquées de l'École Polytechnique de Montréal pour une partie de mon soutien financier.

Table des Matières

Figures

Tableaux

Chapitre 1

Introduction

Les méthodes et algorithmes de classification ou de typologie, dont la science s'appelle la taxinomie, consistent à découper une population en plusieurs classes, en tenant compte des variables qui les caractérisent et d'une mesure de ressemblance entre les objets. La population forme l'ensemble E dont les n éléments sont les sujets ou unités statistiques. Elle est en général décrite par une matrice *sujets* $\times$ *variables*: X, à n lignes et p colonnes. Ainsi chacun des n sujets est décrit par p variables qui sont ou qualitatives (catégoriques) ou numériques.

À l'occasion de quelques rappels de notions de taxinomie, nous mentionnons les difficultés qui confrontent les utilisateurs, et nous exposons la question particulière qui fait l'objet de ce mémoire: soit le problème de consensus entre partitions de E (chapitre 1), dont la solution de Celeux (chapitre 2) est étudiée suivant une approche nouvelle, semble-t-il, faisant appel aux plans d'expériences qu'on trouve le plus souvent maintenant en contrôle de qualité (chapitre 3). Nous terminerons

par une discussion.

1.1 Méthodes hiérarchiques

La taxinomie a vu le jour dans les domaines de biologie, de zoologie, et dans des études de géologie. À l'image des célèbres classifications en espèces, genres, familles, et ordres, le résultat des méthodes hiérarchiques est souvent un diagramme en deux dimensions appelé *dendrogramme*. Celui-ci illustre une suite de partitions emboîtées, partant de la partition des singletons, se terminant par la partition grossière (ou inversement). L'indice de fusion (méthodes ascendantes) ou de séparation (méthodes descendantes) que l'on fait apparaître sur le dendrogramme permet de préciser à quel niveau de l'indice se forment les groupes. On obtient des *hiérarchies indicées*. L'ouvrage de Chandon et Pinson (1981) traite des différentes méthodes hiérarchiques, des choix des indices de dissimilarité et d'agrégation.

On a beaucoup étudié les propriétés des diverses méthodes, et la question de leur validation vient naturellement à l'esprit. Ces méthodes ne sont pas facilement applicables aux grands ensembles de données, vu la mémoire nécessaire à la construction des dendrogrammes. Les méthodes non hiérarchiques prennent le relais dans ce cas (Chandon & Pinson, p.160).

1.2 Méthodes non hiérarchiques

Ces méthodes, appelées aussi méthodes de partitionnement visent à construire une partition des sujets en k groupes, où k est soit fixé *a priori*, soit déterminé par un algorithme. Elles comprennent généralement deux étapes : choix d'une configuration initiale d'une part, et d'autre part l'affectation des éléments aux classes.

La configuration initiale consiste à choisir un certain nombre de noyaux, ou de classes pour amorcer la méthode.

L'affectation, elle, se fait toujours selon un critère qui mesure la distance entre deux classes, laquelle découle d'une mesure d'éloignement entre deux sujets ou unités statistiques. Les différents critères à optimiser découlent du théorème de Huygens généralisé à l'inertie totale T du nuage de points est égale à la somme des inerties inter-classes B et intra-classes W: $T = B + W$. L'optimisation vise soit à minimiser la dispersion au sein de chaque groupe ou, de façon équivalente, à maximiser la séparation entre les classes, puisque l'inertie totale du nuage est constante.

1.3 La méthode des nuées dynamiques

Parmi les méthodes de partitionnement les plus populaires, on trouve la méthode des nuées dynamiques qui généralise la méthode, très naturelle, des centres mobiles. Cette classe de méthodes fut définie par Edwin Diday, et on en trouve

de nombreux développements et applications dans Diday *et al.* (1980). Il s'agit en gros d'optimiser un critère qui exprime l'adéquation entre une classification, ou mode de recouvrement, des sujets et un mode de représentation des classes de cette classification. On peut décrire cette méthode par les quatre items suivants:

a- un mode de représentation:

Un groupe de données est représenté par un noyau qui peut être son centre de gravité, une droite caractérisante, un sous-ensemble de points privilégiés par rapport à d'autres, *etc.* L'ensemble des noyaux constitue l'espace de représentation noté $\mathcal{L}$. On choisit en plus une mesure de distance D d'une unité statistique à un noyau. L'ensemble des données X est muni alors d'une structure de représentation.

b- un espace de recouvrement:

Généralement, on choisit l'ensemble des partitions des sujets en k classes, noté $\mathcal{P}_k$.

c- un critère à optimiser:

C'est une fonction W de $\mathcal{P}_k \times \mathcal{L}_k$ dans $\mathcal{R}_+$. qui mesure l'adéquation entre un recouvrement P et un élément L du produit k fois avec lui même de $\mathcal{L}$, noté $\mathcal{L}_k$.

d- un algorithme :

Il faut définir un algorithme de recherche de cet optimum et examiner ses propriétés de convergence. Cet algorithme repose sur deux fonctions :

- une fonction de représentation: $f : \mathcal{P}_k \longrightarrow \mathcal{L}_k$

- une fonction d'affectation: $g : \mathcal{L}_k \longrightarrow \mathcal{P}_k$

L'algorithme consiste à itérer $f \circ g$ (respectivement $g \circ f$) après l'avoir initialisé à l'aide d'une représentation estimée ***a priori*** ou tirée au hasard (respectivement un recouvrement), et fonctionne en itérant la procédure et en optimisant le critère choisi à chaque étape du déroulement. On montre, sous des conditions générales, que l'on a une convergence vers un recouvrement et une représentation stable en un nombre fini d'itérations (Diday *et al.*, p.119).

1.3.1 Les variantes des nuées dynamiques.

Les variantes de cette méthodes découlent des différents choix de D, de W, de l'espace de représentation $\mathcal{L}$ et de celui de recouvrement. Plusieurs auteurs ont travaillé à appliquer cet algorithme de nature très rapide, et donc propre à traiter de grands ensembles de données, dans leur domaine (*e.g.* P. Rousseau et D. Sankoff, 1979; A. Morin, 1984). Le cas le plus simple des nuées dynamiques est la méthode des centres mobiles, où les noyaux sont les centres des gravité des classes, et la distance entre les sujets la distance euclidienne. Naturellement X est alors une matrice de variables quantitatives. Govært (1983) a développé plusieurs algorithmes de type nuées dynamiques dans le cas de variables catégoriques.

1.4 Problèmes de la taxinomie numérique

Aucune méthode de la taxinomie numérique n'est à l'abri de critiques fondamentales sur ses applications. Les méthodes de classification ne proposent pas sou-

vent (sauf dans les données très bien séparées) une séparation naturelle au "bon" nombre de groupes présents dans les données. Et puis chaque algorithme hiérarchique propose un dendrogramme propre: lequel est le "bon"? Les méthodes de partitionnement, elles, fixent un nombre de groupes *a priori*, imposent en somme une structure aux données qui n'en ont pas nécessairement.

1.4.1 Le bon nombre de groupes

Aucune technique de classification n'apporte beaucoup d'information à ce sujet. Un algorithme prouvera toujours l'existence de classes dans un ensemble de données, même si celui-ci ne jouit d'aucune structure. Un problème déjà à la base: qu'entendre par une absence de structure?

Plusieurs auteurs anglo-saxons développent leurs recherches sur les problèmes relatifs aux tests de présence de structure, (Everitt, 1980). On trouve ainsi des tests de comparaisons: hypothèse de présence de k_1 versus k_2 groupes. On considère beaucoup aussi l'étude du comportement des algorithmes en fonction du nombre de classes présentes dans les données et de celles inférées par une technique ou, plus simplement, par l'utilisateur.

Devant la prolifération considérable des algorithmes de classification, des procédures de simulation, en vue de validation, ne peuvent qu'être d'un bon apport d'information à ce sujet. En fait, la simulation apparaît comme la seule approche à cette question.

De nombreux travaux traitent de ce problème Voir entre autres : Blashfield

(1976), Edelbrock (1979), Dubes et Jain (1979), Bayne *et al.* (1980). Parmi les travaux les plus récents, il faut citer un des plus élaborés, soit un article de Milligan et Cooper (1985). Dans cette importante étude, ces deux auteurs comparent trente(!) critères d'arrêt pour estimer le bon nombre de classes, utilisant quatre méthodes hiérarchiques. Ils utilisent pour ce faire des données simulées qui tentent de se conformer le plus possible à ce qu'ils conçoivent être des situations réelles (Milligan, 1985). Trois facteurs sont considérés: le nombre de groupes (bien séparés), la dimension de représentation (*i.e.* p), et les proportions respectives des groupes, ce qu'ils appellent de façon un peu abusive, la densité. Mais à cause des limites de calcul (ils étudient les méthodes hiérarchiques...), leurs ensembles n'ont que cinquante points, en dimension jusqu'à huit!

Bien évidemment, on a songé à utiliser de nouvelles techniques d'estimation appliquées à ce problème. Citons ici, pour mémoire, Jain et Moreau (1986) qui utilisent une procédure Bootstrap et le critère de Davies et Bouldin pour mesurer la stabilité d'une partition.

Nous décrirons avec plus de précision nos choix au chapitre trois.

1.4.2 La notion de concensus

Depuis longtemps les taxinomistes s'intéressent au problème de concensus entre les classifications. On pourra consulter à ce sujet le numéro spécial du "Journal or Classification" (volume 3 n°2, 1986) consacré à cette question. Ceci se justifie par la présence d'une multitude d'algorithmes, qui généralement s'appuient sur des critères générant des solutions optimales localement. Ainsi, les nuées dynamiques,

quelle que soit la variante, demandent une initialisation et donc les solutions en dépendent.

La situation générale dans lesquelles se présente la question de concensus est la suivante: Soient $M_1, M_2, \ldots, M_v, v$ modèles de classification d'un groupe de sujets, v partitions donc. Comment déterminer un modèle qui rende compte au mieux des v modèles? Comment synthétiser l'information apportée par chaque classification afin de définir un *consensus*?

Leclerc et Cucumel (1988) présentent une bibliographie commentée du problème des consensus.

On distingue communément trois approches pour aborder cette question: par optimisation, par construction, ou par l'approche axiomatique. On trouve une synthèse des ces procédures faite par Barthélemy, Leclerc et Monjardet (1984).

Dans l'approche par optimisation, on évoque souvent les partitions centrales. On se fixe alors un critère basé sur une métrique d dans l'ensemble M des modèles, laquelle mesure l'éloignement d'un modèle par rapport à un autre.

Soient $M = \mathcal{P}(X)$, l'ensemble des partitions de X, $\mathcal{RE}(X)$ l'ensemble des relations d'équivalence associées, d une distance entre deux éléments de M, par exemple la distance de la différence symétrique des graphes des relations d'équivalence associées.

Si $P_1, P_2, \ldots, P_v$ sont v partitions de X, associées respectivement aux relations d'équivalence $u_1, u_2, \ldots, u_v$, une partition centrale P_0 est celle dont la relation

d'équivalence associée, minimise la fonction suivante:

$$D(u) = v^{-1} \sum_{i=1}^{v} d(u, u_i)$$

où le minimum est pris sur l'ensemble $\mathcal{RE}(X)$.

Diverses heuristiques ont été élaborées à ce sujet (*e.g.* Marcotorchino et Michaud, 1982; Régnier, 1983). Celeux (1984) pour sa part présente une approximation rapide du calcul de la partition centrale.

L'objet de ce mémoire est une validation d'une heuristique de Celeux dans le cas où on utilise les centres mobiles comme technique de partionnement. Nous voulons utiliser, pour ce faire, des idées tirées du contrôle de la qualité.

Au chapitre suivant, nous décrivons cette heuristique, puis, au chapitre trois, les plans d'expérience retenus, ainsi que les conclusions auxquelles nous parviendrons.

Chapitre 2

L'heuristique de Celeux

Dans ce chapitre, nous énonçons précisément l'objet de notre étude: l'heuristique de Celeux appliquée aux multipartitions issues de la méthode des centres mobiles.

2.1 La méthode des centres mobiles

Cette méthode de classification est la variante paradigmatique des nuées dynamiques. X est une matrice de données où les variables sont quantitatives, et où la distance euclidienne entre les variables définit la métrique entre les sujets qui forment l'ensemble E. Ainsi un sujet peut être identifié à un vecteur de $\mathcal{R}^p$. Si X_i et X_j sont deux sujets, ou deux lignes de X, leur distance est définie par l'équation:

$$d^2(X_i, X_j) = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$$

L'espace de recouvrement de E est P_k, l'ensemble des partitions de E en k sous-ensembles. L'ensemble des représentations est $\mathcal{L} = \mathcal{R}^p$.

Chaque groupe, ou sous-ensemble d'une partition, est représenté par son centre de gravité. La distance d'une unité statistique à un noyau est encore la distance euclidienne. Celle d'un sous-ensemble d'une partition à son noyau est, bien entendu, la somme des distances de chacun des sujets du sous-ensemble au noyau. Une partition de l'espace de recouvrement est représenté par le k-uplet des centres de gravité, un élément donc de $\mathcal{L}_k = \mathcal{R}^p \times \cdots \mathcal{R}^p$, k fois.

La fonction W à optimiser est l'inertie intra-classe d'une partition. On en cherche le minimum. On la définit par les relations suivantes:

$$W = \sum_{c=1}^{k} W_c$$

avec $W_c = \sum_{i=1}^{n_c} d^2(X_i, G_c)$ où n_c est la cardinalité de la classe c, dont G_c est le centre de masse.

Restent à définir les deux fonctions, f de représentation et g d'affection dont les itérées, $g \circ f$ (resp. $f \circ g$), vont converger vers la meilleure représentation, étant donnée une représentation initiale (resp. un recouvrement initial).

Dans notre cas, nous initialisons l'algorithme avec un k-uplet de représentants.

À un k-uplet de centres amorces, $G_0 = (g_1^0, \ldots, g_k^0) \epsilon \mathcal{L}_k$, la fonction d'affectation

g assigne la partition des sujets $P_j, j = 1, \ldots, k$, où

$$P_j = \{x \epsilon E : d(x, g_j^0) < d(x, g_i^0), i \neq j\}$$

Le représentant dans $\mathcal{L}_k$ associé à une partition est constitué du k-uplet des centres de gravité de chacun des éléments de la partition.

L'algorithme qui itère successivement les fonctions d'affectation et de représentation, à partir d'une représentation initiale, fait décroître l'inertie intra-classes. On montre également que la convergence se fait en un nombre fini d'itérations (Diday *et al.*, 1986).

Géométriquement, on conçoit que cet algorithme a tendance à déterminer des classes bien "rondes". De même, les surfaces séparatrices de toute paire de classes sont des hyperplans perpendiculaires aux segments des centres de gravité.

L'inconvénient majeur de cette méthode est, ainsi qu'on l'a déjà mentionné, sa vulnérabilité à l'initialisation. Des initialisations différentes vont en général donner des optimums différents, et des partitions associées distinctes. On n'est assuré en fait que d'un minimum local au critère à optimiser W, soit l'inertie intra-classes.

Pour pallier à cet handicap, on peut recommencer m fois la procédure en partant de différents choix initiaux, obtenant ainsi ce qu'on appelle une multipartition de E, de longueur m. On peut s'intéresse alors à une ***superposition*** des solutions: les individus classés ensemble dans cette nouvelle partition sont ceux qui se retrouvent ensemble lors des m partionnements de E. Ces sous-ensembles de nature stable, on les nomme les *formes fortes*, et leur réunion constitue une partition de E. Cependant celle-ci est en général trop fine pour répondre aux besoins de l'uti-

lisateur. On y trouve le plus souvent en effet un très grand nombre de classes, et leur interprétation est le plus souvent très fragile, sinon impossible. On verra plus loin une utilisation des formes fortes pour réduire certains calculs.

2.2 Partitions centrales

La notion de partition centrale est une tentative pour définir par optimisation une notion de concensus entre plusieurs partitions.

Soient P_1, P_2 deux partitions de E, auxquelles sont associées respectivement les relations d'équivalence u_1, u_2. On mesure l'éloignement entre ces deux partitions par la métrique suivante : si Gu_i est le graphe de la relation u_i, alors

$$d(P_1, P_2) = |Gu_1 \triangle Gu_2|$$

où $|.|$ dénote le cardinal d'un ensemble, et $\triangle$ est l'opérateur de différence symétrique entre deux ensembles.

On peut définir, pour une relation d'équivalence u sur E, une fonction notée par la même lettre sur $E \times E$ dans $\{0,1\} \subset \mathcal{R}$ ainsi définie:

$$u(x,y) = \begin{cases} 1 & \text{si } x \text{ est en relation avec } y \\ 0 & \text{sinon} \end{cases}$$

e. la fonction caractéristique du graphe de la relation dans E^2.

Supposons qu'on a $\{P_i\}, 1 \leq i \leq m$, m partitions de E; les partitions centrales

P sont celles qui minimisent $\sum_{i=1}^{m} d(P, P_i)$. Soit maintenant la fonction numérique s définie sur $E \times E$ par:

$$s(x,y) = m^{-1} \sum_{i=1}^{m} u_i(x,y),$$

c'est à dire la fonction: moyenne des fonctions caractéristiques des relations d'équivalence telles que définies plus haut. Lerman (1970) montre qu'une partition centrale de relation d'équivalence associée $u(x,y)$, maximise la forme linéaire suivante:

$$L(u) = \sum_{(x,y)\epsilon E^2} (u(x,y) - 1/2)(s(x,y) - 1/2)$$

Ce critère est très important pour notre étude. Il en découlera certaines conséquences sur la nature des éléments de la multipartition dans le cas où E ne jouit d'aucune structure.

2.3 L' heuristique

Il est naturellement impossible d'effectuer une recherche exhaustive sur toutes les relations d'équivalence sur E pour trouver le minimum du critère. Ce problème est d'ailleurs NP-complet (Celeux, 1984). Diverses solutions ont été proposées ainsi que montionné plus haut. Pour notre part nous examinerons une heuristique que Gilles Celeux a décrite dans un rapport de l'INRIA où il en montre également l'utilisation sur quelques exemples de données réelles (Celeux, 1984).

Soient $\{P_i : 1 \leq i \leq m,\ P_i \epsilon \mathcal{P}(E)\}$, une multipartition de E, et soient u_i les

relations d'équivalence associées. Pour tout $u \epsilon \mathcal{RE}(E)$, la distance entre u et la multipartition est donnée par:

$$C(u) = \sum_{(x,y)\epsilon E^2} \sum_{i=1}^{m} |u(x,y) - u_i(x,y)|$$

On recherche $\min C(u)$ sur l'ensemble $\mathcal{RE}(E)$ des relations d'équivalence sur E. Définissons maintenant la matrice $n \times n$, $A = (a(x,y))$:

$$\text{pour } (x,y)\epsilon E^2,\ a(x,y) = \textstyle\sum_{i=1}^{m} u_i(x,y).$$

À partir de cette matrice A, on construit $m+1$ relations binaires sur E:

pour

$$0 \leq l \leq m,\ v_l(x,y) = \begin{cases} 1 & \text{si } a(x,y) \geq l \\ 0 & \text{sinon} \end{cases}$$

Du fait que les u_i sont des relations d'équivalence, les v_l sont symétriques et réflexives. Mais seules v_0 et v_m sont nécessairement transitives. Elles correspondent respectivement à la partition grossière et la partition des formes fortes des P_i.

Soient $\overline{v}_l$, $0 \leq l \leq m$, leurs fermetures transitives. Deux restrictions sont considérées par Celeux dans le problème de la partition centrale:

- La première est de chercher le minimum de $C(u)$ dans le sous-ensemble de $\mathcal{RE}(E)$ comprenant seulement les $\overline{v}_l$, $0 \leq l \leq m$. On obtient ainsi une relation d'équivalence sous-optimale pour le critère. On l'appelera la partition *pseudo-centrale.*

- La seconde est de ne plus travailler sur E, mais sur le sous-ensemble F des représentants des formes fortes, *i.e.* l'ensemble quotient de E par la relation des formes fortes. Ceci a pour effet de réduire grandement les calculs, puisque la cardinalité de F est en général de beaucoup plus petite que celle de E.

On peut montrer (Celeux, 1984) que le critère $C(u)$ se réduit alors à trouver le minimum pour $l = 0, \ldots, m$ de:

$$C(\overline{v}_l) = 1/2 \sum_{i=1}^{m} \sum_{(f,f') \epsilon F^2} (\text{card} f)(\text{card} f') | \overline{v}_l(f,f') - u_i(f,f') |$$

où card$f$ est le nombre d'éléments de E dans la forme forte représentée par f.

Celeux fait valoir que cette méthode retrouve impeccablement la structure de quelques ensembles bien typés, et qu'elle ne donne pas de résultats clairs là où la structure des données n'est pas du tout nette.

La procédure de détermination du "bon" nombre de classes que Celeux propose consiste à faire varier k, le nombre de groupe requis pour la méthode des centres mobiles, dans une certaine plage à l'intérieur de laquelle on espère trouver la vraie valeur. Pour les ensembles bien typés, la partition pseudo-centrale obtenue donne un nombre de classes constant sur plusieurs k, et cette constance est interprétée comme une indication du bon nombre de classes. Sur les données réelles qu'il utilise on trouve bien le bon nombre.

Il semble bien qu'on ait là une tentative intéressante pour se libérer de l'influence de l'initialisation des centres mobiles. Mais restent ouvertes les questions de la validation de cette heuristique. Ce n'est pas sur quelques exemples qu'on

peut juger. C'est une chose de retrouver une structure connue, c'en est une autre de trouver correctement une structure inconnue, ou encore d'en donner une interprétation valable.

Au chapitre suivant, nous décrivons les choix que nous avons faits des facteurs qui peuvent influencer la décision sur le nombre de groupes présents dans les données, de même que l'expérience statistique que nous avons faite. Nous verrons alors surgir un certain nombre de questions qui n'ont pas encore de réponses satisfaisantes, et qui pourront faire l'objet de prochaines études.

Remarque terminologique: Dans la suite de ce travail, nous appelons *heuristique de Celeux* la suite de procédures comprenant l'algorithme des centres mobiles suivi de la détermination de la partition pseudo-centrale. Cela suppose la fixation du nombre de points amorces, le paramètre k, et la longueur de la multipartition m.

Chapitre 3

Les expériences

3.1 Introduction

Une expérience est une intervention volontaire
dans un système pour observer ou mesurer
les effets de cette intervention.
Michel Vigier.

Mener un plan d'expérience, c'est d'abord identifier les facteurs qui affectent une certaine variable à laquelle on s'intéresse, et mesurer les effets produits sur cette variable si on perturbe les facteurs. Une forme de perturbation consiste à faire passer les facteurs d'une modalité à une autre.

Dans un plan d'expérience orthogonal, chacun des effets est mesuré de sorte que les autres n'apportent aucun biais à son estimation. Nous ne considérerons ici que des expériences où les facteurs sont à deux modalités.

On estime qu'une planification d'expérience est rigoureusement adéquate pour nous fournir les résultats espérés vu les outils d'analyse qu'elle met à notre disposition. Elle nous permettera d'identifier certains facteurs qui affectent la tendance centrale et la variabilité de la performance de l'heuristique de Celeux.

Nous l'avons mentionné plus haut, nous ne voulons pas simuler des situations réelles. Semblable en cela par exemple à la mécanique des fluides, où les méthodes de calcul sont testés sur des géométries très simples, standardisées en quelque sorte, notre objectif est de définir des conditions expérimentales extrêmement simples. Elles pourront ainsi servir de base de comparaison pour tester des raffinements éventuels de l'heuristique de Celeux, ou pour d'autres algorithmes.

3.2 Le premier plan

Il est bien évident que nous devons limiter l'envergure de notre étude. On ne pourra donc examiner tous les facteurs qui viennent à l'esprit lorsqu'on réfléchit aux problèmes méthodologiques de la classification automatique.

Nous ne travaillons ici qu'en dimension deux, le plan euclidien usuel ($p = 2$). De plus nos classes auront toutes le même nombre de points.

Les facteurs considérés sont de deux catégories : les facteurs propres aux données générées, et ceux propres à l'heuristique de Celeux.

3.2.1 Les facteurs propres aux données générées

Dans une première expérience où l'on s'intéresse à une "absence" de structure, on considère les deux répartitions les plus propres à simuler cette situation: la distribution binormale centrée avec matrice de covariance définie par la matrice identité, et la distribution uniforme sur le disque unité. On aurait pu considérer aussi, en dimension deux, un produit de lois uniformes. On retrouve ces hypothèses par exemple dans Bock (1979) ou encore dans Lerman (1970).

Dans nos expériences, nous considérerons la géométrie la plus simple possible: les groupes seront répartis sur un polygone régulier, et nous utiliserons un indice de séparation des groupes qui est la distance au centre de masse des sujets des sommets du polygone régulier. Il faut mentionner que dès qu'il y a plus de trois groupes, les distances entre les paires de groupes ne sont pas constantes.

Les deux séparations considérées sont de quatre et de huit, sauf dans le cas de la première expérience où il y a ou deux groupes, à distance quatre l'un de l'autre, ou un seul groupe. Dans les deux cas les groupes sont très fortement séparés. Chaque groupe possède cinquante sujets, sauf dans le cas d'absence de structure où le groupe unique est de cardinalité cent.

3.2.2 Les facteurs propres au programme

- *Le nombre des classes:* C'est le paramètre k qui doit être fixé pour amorcer l'algorithme des centres mobiles. Les niveaux fixés sont 3 ainsi que 7.

Dans l'article de Celeux, on utilise la technique qui consiste à faire varier

séquentiellement k. Dans le contexte des expériences où les facteurs ont deux niveaux, nous ne pouvons respecter cette procédure, mais des essais préliminaires nous ont induit à ne considérer que cette procédure où les deux modalités sont assez extrêmes l'une par rapport à l'autre.

- *La longueur de la multipartition:* C'est le nombre m de fois qu'on doit initialiser l'algorithme de partionnement afin de former une multipartition. Nous avons choisi 5 et 15. Encore une fois des valeurs extrêmes.

- *La nature des points amorces:* Nous avons retenu le choix au hasard des noyaux initiaux, ainsi que le choix de noyaux alignés.

On trouve de nombreuses variantes pour les choix initiaux (Anderberg, 1973), la plus populaire étant l'amorce sur des sujets choisis au hasard. Mais nous avons voulu tester une variante originale: choisir des noyaux alignés.

Le choix des noyaux alignés se fait de la façon suivante: deux sujets sont choisis au hasard, puis les noyaux suivants sont choisis au hasard sur la droite joignant les deux initiaux. On note que ces derniers ne correspondent pas souvent à des sujets, mais cela n'a aucune importance puisque dès la deuxième itération de l'algorithme des centres mobiles, les noyaux ne sont plus en général des sujets de la population observée.

3.2.3 La variable dépendante

Nous devons définir une réponse qui mesure la performance de l'heuristique. Pour une certaine combinaison des facteurs cités ci-dessus, on obtient une parti-

tion pseudo-centrale, qui sera comparée à la structure originale. On attribue une variable logique qui vaut 1 si le critère retrouve la structure simulée, 0 sinon. On recommence cent fois l'heuristique dans ces conditions et la variable dépendante considérée est le pourcentage des fois où la variable logique vaut 1.

Dans chaque expérience, nous avons répété trois fois chaque essai.

3.2.4 Les effets et résolution

L'effet d'un facteur est la différence entre les moyennes des réponses pour lesquelles le facteur est au niveau le plus haut et celle pour lesquelles celui-ci est au niveau le plus bas. L'effet, sur la variable réponse, de chacun des facteurs cités ci-dessus, est dit principal. Un facteur d'ordre p — ne pas confondre ici avec le nombre de variables descriptives des sujets —, désigne une interaction de p facteurs, dont on mesure l'effet du même ordre sur la réponse.

Un schéma d'expérience est dit de résolution R si tout effet d'ordre p n'est confondu avec un effet contenant moins de $R - p$ facteurs. Par exemple, en résolution III, aucun effet principal n'est confondu avec un effet principal. En fait, la résolution d'un schéma à deux niveaux est la longueur du plus court mot dans la relation de définition.

Dans nos expériences, les schémas que nous avons considérés sont de résolution V.

3.3 Première expérience

3.3.0.1 Résumé des facteurs

Au tableau suivant, on trouve le résumé des facteurs de notre première expérience.

Avec 5 facteurs, chacun à deux modalités, on aurait normalement 32 combinaisons à considérer. Seulement une fraction de ce nombre peut nous fournir aussi l'information visée à travers le schéma complet. On sera ainsi amené à choisir un schéma fractionnaire noté 2^{5-p}, où p est le nombre de générateurs .

On définit le cinquième facteur par l'interaction d'ordre quatre notée 1234, c'est le générateur.

Le nombre d'essais est alors fixé à 16, et la relation de définiton du plan d'expérience possède seulement deux mots : I=12345, où I désigne la colonne avec des +.

La liste des confusions dérive automatiquement de la relation I=12345, en considérant que le produit de chaque facteur par lui-même est égal à I. L'égalité 5=1234 signifie que l'effet du cinquième facteur est confondu avec celui de l'interaction d'ordre quatre 1234. La liste complète des confusions se trouve au tableau 3.2 de même que le schéma de la première expérience.

facteur	signification	modalités
1	longueur de la multipartition.	+:15
		−: 5
2	nombre de points amorces.	+: 7
		−: 3
3	nature des points amorces.	+:alignés
		−:au hasard
4	nature des données générées.	+: Unif.
		−: Gaus.
5	séparation entre les classes.	+: 4
		−: 0

Tableau 3.1: Résumé des facteurs.

3.3.1 Analyse

- On a k objets à comparer, c'est-à-dire, k moyennes $\mu_1, \ldots, \mu_k$. Ici $k = 16$. On appelle contraste une combinaison linéaire des moyennes μ_j telle que la somme des coefficients est nulle: $C = \sum_{j=1}^{k} C_j \mu_j$, avec $\sum C_j = 0$. On prend les C_j égaux à ± 1, suivant le signe de la modalité considérée.

- Dans le cas d'un plan d'expérience orthogonal, on a $k-1$ contrastes orthogonaux, et la variation inter-groupes se décompose en somme des carrés des estimations des $k-1$ contrastes.

$$\mathrm{SSB} = \sum_{j=1}^{k-1} \mathrm{SS}(C_j)$$

Un contraste est estimé par $\hat{C}$:

1 = 2345	2 = 1345
3 = 1245	4 = 1235
5 = 1234	12 = 345
13 = 245	14 = 235
15 = 234	23 = 145
24 = 135	25 = 134
34 = 125	35 = 124
45 = 123	

	facteurs				
essai	1	2	3	4	5
1	-	-	-	-	+
2	+	-	-	-	-
3	-	+	-	-	-
4	+	+	-	-	+
5	-	-	+	-	-
6	+	-	+	-	+
7	-	+	+	-	+
8	+	+	+	-	-
9	-	-	-	+	-
10	+	-	-	+	+
11	-	+	-	+	+
12	+	+	-	+	-
13	-	-	+	+	+
14	+	-	+	+	-
15	-	+	+	+	-
16	+	+	+	+	+

Tableau 3.2: Les confusions et le schéma de la première expérience.

$$\hat{C} = \sum_{j=1}^{k} C_j \overline{y}_j$$

où $\overline{y}_j$ est la moyenne des 3 réponses à l'essai j.

- La somme des carrés associés au contraste C est notée $\mathrm{SS}(\hat{C})$:

$$\mathrm{SS}(\hat{C}) = \hat{C}^2 / n^{-1} \sum_{j=1}^{k} C_j^2 = \hat{C}^2 / n^{-1} 16,$$

et $n = 3$ est égal au nombre de réplications. Chaque contraste est à un degré de liberté, donc la moyenne des sommes des carrés associée à chaque contraste, $\mathrm{MS}(\hat{C})$, est égale à $\mathrm{SS}(\hat{C})$.

- Le ratio-F associé au contraste C, $\mathrm{F}(\hat{C}) = \mathrm{SS}(\hat{C})/\mathrm{MSW}$, où MSW est la variation intra-groupes moyenne, SSW divisée par son degré de liberté.

- L'effet associé au contraste C, noté $\hat{l}$:

$$\hat{l} = \hat{C} / \sum_{C_j > 0} C_j.$$

$\hat{l}$ estime la différence entre les moyennes des groupes comparés par le contraste C.

- signification d'un contraste: un contraste C est significatif au niveau α si: $\mathrm{F}(\hat{C}) > \mathrm{F}_{1,\nu,\alpha}$, où ν est le degré de liberté de la variation intra-groupes: $\nu = k(n-1) = 16 \times 2 = 32$.

On trouve au tableau 3.3 les résultats des seize essais.

essais	y_1	y_2	y_3	$\overline{y}$	s^2
1	.46	.38	.47	.436	.002433
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0
6	.99	.97	1	.986	.000233
7	.45	.62	.32	.463	.02263
8	0	0	0	0	0
9	0	0	0	0	0
10	.69	.83	.78	.76	.00503
11	0	0	0	0	0
12	0	0	0	0	0
13	.33	.58	.38	.43	.018
14	0	0	0	0	0
15	0	0	0	0	0
16	.78	.92	.73	.81	.0097

Tableau 3.3: Résultats de la première expérience.

SOURCE	SS	DF	MS	F
Inter	5.580	15	.372	103.33
Intra	.115	32	.0036	
Totale	5.695	47		

Tableau 3.4: Analyse de la variance pour nos données.

On trouve maintenant: $\overline{s^2} = (1/16)\sum_{j=1}^{16} s_j^2 = .0036$, ceci estime la variation intra-groupes moyenne. Au tableau 3.4, on donne le tableau d'analyse de variance de l'expérience.

Du fait que le ratio F= MSB/MSW = 103.33 suit une distribution de Fisher à 15 et 32 degrés de liberté, la différence entre les seize groupes est très significative. Le but de notre analyse est loin d'être atteint, puisque ce qui nous intéresse est de connaître les effets de chacun des facteurs sur la variation totale (SST).

On trouve finalement au tableau 3.4 les résultats des calculs de signification pour les facteurs et leurs interactions.

Comme au seuil de 5%, on a $F_{1,32,.05} = 4.15$, on retient à ce seuil les facteurs 1,2,3 et 5. Tous les facteurs sont très significatifs sauf le quatrième qui est la nature de la distribution. Le plus étonnant, c'est que ce dernier interagit fortement avec la plupart des autres facteurs. Le test de Fisher est très robuste dans le cas où les hypothèses de normalité sont respectées. Faut-il le retenir ou l'écarter?

On a préféré jeter un regard sur le taux de variation expliquée par les facteurs. Si on retenait dans le modèle les facteurs 2, 3 et 5 avec leurs interactions,le taux d'explication est de l'ordre de 78%. En ajoutant le facteur 1, le modèle explique

Facteurs & interactions	$\hat{I}$	$\hat{C}$	SS($\hat{C}$)	F($\hat{C}$)	% variation
1	0.154	1.227	0.282	78.3	4.95
2	-0.168	-1.366	0.350	97.2	6.1
3	0.186	1.493	0.418	116	7.3
4	0.015	0.115	0.002	0.55	0.035
5	0.490	3.885	2.832	786.7	49.7
12	-0.070	-0.533	0.054	15	0.9
13	0.072	0.579	0.063	17.5	1.1
14	0.131	1.053	0.209	58.05	3.7
15	0.153	1.227	0.282	78.33	4.9
23	0.132	1.053	0.208	57.77	3.6
24	0.072	0.579	0.063	17.5	1.1
25	-0.167	-1.339	0.336	93.33	5.9
34	-0.067	-0.533	0.053	14.72	9
35	0.186	1.493	0.418	116.11	7.3
45	0.015	0.115	0.002	0.55	0.05

Tableau 3.5: Calculs de signification des facteurs et pourcentages de variation expliquée.

près de 92% de la variation totale. En fait, le quatrième facteur n'explique que .03% ce qui est très négligeable. Mais, même s'il n'est pas statistiquement significatif, il interagit assez fortement, on l'a déjà noté, avec les autres facteurs! On a décidé toutefois de ne pas le retenir dans les autres plans d'expérience. En partie pour la raison qu'on vient d'exposer, et en partie à cause du diagramme de Daniel qu'on présente maintenant.

La méthode de Daniel est basée sur l'utilisation du papier gaussien. Supposons que les données sont le résultat de variations aléatoires autour d'une moyenne fixe et que les changements de niveaux des facteurs n'ont pas d'effet réel sur la réponse. Alors les effets, qui sont des différences de moyennes de groupes des observations, sont distribués suivant une loi gaussienne. On porte alors les effets sur du papier gaussien, et les points non alignés ne vérifient pas l'hypothèse de ne pas avoir d'effet. On voit sur la figure 3.1 que les facteurs 1, 2, 3 et 5 ainsi que leurs interactions ont un effet sur les résultats de l'expérience. Les autres peuvent être expliqués par le bruit.

En résumé, on retient au niveau 5% les facteurs 5, 3, 2 et 1 (mentionnés par ordre d'importance), ainsi que leurs interactions.

Le diagramme de Pareto (figure 3.2) illustrant les pourcentages de variation expliquée par chacun des facteurs, vient confirmer les conclusions obtenues par la méthode de Daniel.

Notons maintenant, à partir du tableau 3.5, que le choix de noyaux de départ alignés donne de meilleurs résultats que le choix au hasard. Un choix de trois noyaux, soit près de la réalité (deux classes), semble donner un meilleur résul-

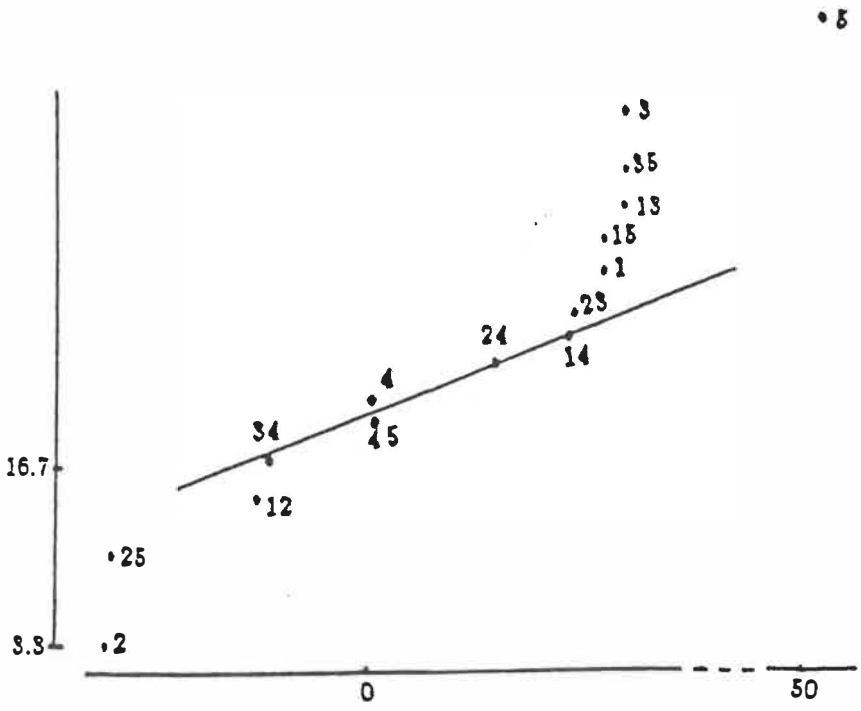

Figure 3.1: Diagramme de Daniel pour la première expérience.

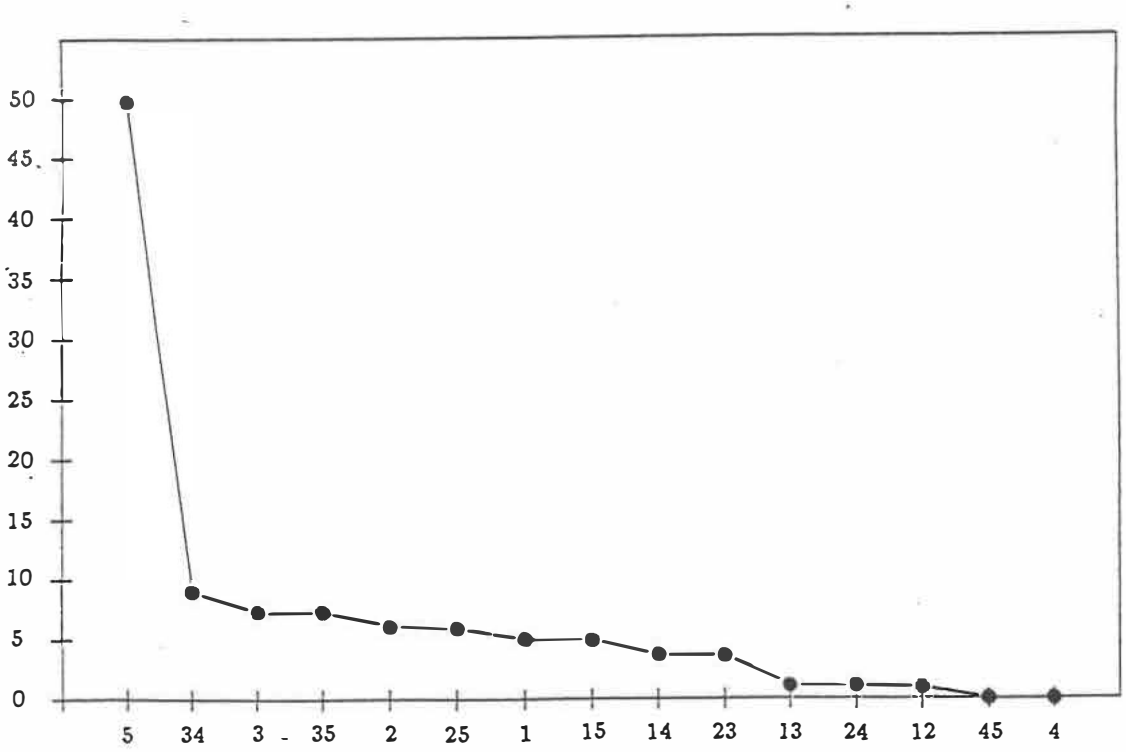

Figure 3.2: Diagramme de Pareto pour la première expérience. En abscisse, les facteurs par ordre décroissant d'importance; en ordonnée les pourcentages de variation expliquée par les facteurs.

tat. On voit bien, par exemple, qu'à l'essai quatre $k = 7$ points amorces dans l'algorithme ne retrouve jamais la présence des deux groupes mais s'ils sont très fortement séparés...

Fait beaucoup plus remarquable et intéressant, cependant, lorsqu'il n'y a pas de structure, *i.e.* lorsqu'on trouve dans les données un seul groupe qu'il soit issu d'une gaussienne centrée réduite ou d'une loi uniforme sur le disque, on ne retrouve jamais une partition pseudo-centrale qui soit grossière comme on serait en droit d'espérer. Ainsi la partition pseudo-centrale impose dans ce cas une structure à des données qui n'en ont pas, exactement comme l'algorithme des centres mobiles lui-même.

En conséquence de quoi, dans les expériences suivantes, jamais on ne considère ce qu'on pourrait appeler l'hypothèse nulle, soit un seul groupe.

La question se pose donc de connaître à quelles conditions une multipartition donnerait la partition grossière comme partition centrale. Une réponse partielle se trouve peut-être dans les remarques suivantes.

3.3.2 Comment obtenir la partition grossière?

Un grand nombre d'essais nous incitent à conjecturer la proposition suivante: Ce n'est que lorsque plus de la moitié des partitions d'une multipartiton sont les partitions grossières qu'on obtient nécessairement celle-ci comme partition optimale. Nous prouvons ici ce résultat dans le cas le plus simple, c'est à dire pour une multipartition de longueur deux.

Dans la section 2.2, nous avons rappelé le critère suivant: la partition centrale doit maximiser la forme linéaire :

$$L(u) = \sum_{(x,y)\epsilon E^2} (u(x,y) - 1)(s(x,y) - \frac{1}{2}) \tag{3.1}$$

où $s(x,y)$ est la moyenne des u_i, chaque u_i étant la fonction caractéristique du graphe de la relation d'équivalence correspondante dans E^2.

Dans le cas où on ne considère que deux relations d'équivalence u_1 et u_2, notons par Gu_i le graphe de u_i, et divisons E^2 en quatre parties F_i:

$$F_1 = Gu_1 \cap Gu_2,\ F_2 = Gu_1 \cap \overline{Gu_2},\ F_3 = \overline{Gu_1} \cap Gu_2,\ F_4 = \overline{Gu_1} \cap \overline{Gu_2}.$$

On a donc $s(x,y) = 1$ sur F_1, $s(x,y) = \frac{1}{2}$ sur F_2 et F_3, et $s(x,y) = 0$ sur F_4. L'équation 3.1 devient alors:

$$L(u) = \frac{1}{2}\left(\sum_{(x,y)\epsilon F_1} (u(x,y) - \frac{1}{2}) - \sum_{(x,y)\epsilon F_4} (u(x,y) - \frac{1}{2}) \right) \tag{3.2}$$

Lemme 1 *Avec les notations précédentes, on a que la partition grossière P_o maximise 3.2 si et seulement si $F_4 = \phi$.*

La suffisance est évidente. Pour la nécessité, supposons que P_o est optimale et que F_4 est non vide.

Considérons alors $u = u_1 \wedge u_2$. $Gu = Gu_1 \cap Gu_2$, et $\overline{Gu} = F_2 \cup F_3 \cup F_4$. Alors, si l'on note par $|.|$ la cardinalité d'un ensemble:

$$L(P_\circ) = \frac{1}{4}|F_1| - \frac{1}{4}|F_4|,$$

et

$$L(u) = \frac{1}{4}|F_1| + \frac{1}{4}|F_4|,$$

de sorte que $L(u) > L(P_\circ)$, ce qui contredit l'hypothèse que $P_\circ$ est optimale.

On est en mesure maintenant de montrer la

Proposition 1 *Soit une multipartition comprenant deux partitions d'un ensemble E dont les relations d'équivalence associées sont u_1, u_2. Alors une condition nécessaire et suffisante pour que la partition grossière $P_\circ$ soit une partition centrale de cette multipartition, est qu'une des deux partitions soit elle même grossière.*

(a) Suffisance. Une des deux partitions étant grossière, il s'en suit que $F_4 = \phi$. Donc $L(u)$ dans l'équation 3.2 ne contient que la première somme, et il est évident que la partition grossière est une de celles qui maximise cette somme.

(b) Nécessité. En vertu de lemme, on doit avoir que $F_4 = \phi$. Supposons maintenant que u_1 n'est pas la partition grossière, *i.e.* que $Gu_1 \neq E^2$. On a par définition de F_4:

$$F_4 = \phi \Longleftrightarrow \overline{Gu_1} \cap \overline{Gu_2} = \phi,$$

d'où, par complémentation:

$$Gu_1 \cup Gu_2 = E^2 \tag{3.3}$$

Soit maintenant $(x, y) \epsilon Gu_1$. Alors x et y sont dans une classe de la relation d'équivalence représentée par a_k, disons, dans l'ensemble quotient par la relation u_1. Ainsi pour les autres représentants, disons a_j, on a que x n'est pas en relation u_1 avec aj, de même pour y et a_j. Mais, puisque $Gu_1 \cup Gu_2 = E^2$, on doit avoir $u_2(x, a_j) = 1 = u_2(y, a_j)$, ce qui implique, par transitivité, que $(x, y) \epsilon Gu_2$.

On a montré que $Gu_1 \subset G_2$, et donc en vertu de (3.3), on a $Gu_2 = E^2$, la relation u_2 est la relation grossière.

Il a été impossible jusqu'à maintenant de généraliser cette argumentation à plus de deux partitions. Mais de nombreuses simulations semblent indiquer que si plus de la moitié des partitions d'une multipartition sont la partition grossière, alors la partition grossière est le résultat de l'heuristique de Celeux, alors que celle-ci ne vient jamais dans le cas contraire. Ces simulations ont été faites dans le cas où l'ensemble de données n'avait aucune structure comme dans le cas où on y trouvait deux groupes très bien séparés. On note évidemment que la proposition est valable quel que soit l'ensemble qui a été partitionné...

En somme, cette heuristique trouve une limitation fondamentale: on ne pourra jamais y trouver comme partition pseudo-centrale la partition grossière, puisque la multipartition provient de m applications de l'algorithme des centres mobiles. De plus ajouter un certain nombre de partitions grossières à la multipartition n'est d'aucune utilité.

La question reste ouverte: comment trouver la partition grossière comme partition centrale seulement dans les cas où elle fait sens?

3.4 La deuxième expérience

De la première expérience, on a conclu que le facteur de distance entre les classes jouait un rôle très important puisque, entre autres, il expliquait 50% de la variation totale. En plus, nous ne pouvons plus considérer l'absence de structure, soit la présence d'un seul groupe. Dans notre deuxième expérience, nous considérons le facteur binaire: trois ou cinq groupes, définis par le triangle et le pentagone réguliers situés sur des cercles du plan. Chaque groupe contient cinquante points, ou sujets, et ils sont distribués selon une loi binormale centrée avec matrice de covariance définie par la matrice identité. La distance au centre de gravité de l'ensemble de chacun des centres de gravité des groupes est soit de quatre soit de huit. Cela définit un second facteur binaire. Nous avons également gardé les facteurs du nombre de partitions dans la multipartition, et des deux types d'initialisation des centres mobiles. Enfin, derniers facteurs retenus: la longueur de la multipartition, et le nombre de groupe requis par les centres mobiles. Au tableau 3.6 on trouve le résumé des facteurs.

Le schéma retenu est un 2^{5-p}, de résolution V (5=1234) répété 3 fois. En outre des effets principaux, les interactions d'ordre deux seront aussi estimées. L'expérience est identique à la précédente à une permutation des facteurs près, et on trouve au tableau 3.7 les essais qui la définissent.

Facteur	Signification	Modalités
1	longueur de la multipartition	- : 5 + : 15
2	nombre de groupes	- : 3 + : 5
3	séparation entre les groupes	- : 4 + : 8
4	nature des points amorces	- :alignés + :hasard
5	nombre de points amorces	- : 3 + : 7

Tableau 3.6: Résumé des facteurs de la seconde expérience.

Du fait que le ratio F= MSB/MSW = 766.7, tel que montré à l'analyse de variance de l'expérience (tableau 3.9) suit une loi de Fisher $F_{15,32}$ dont le 95^e percentile est à 4.15, on a évidemment une grande différence entre les groupes. Le tableau 3.10 nous montre l'importance des effets.

La séparation des classes, le nombre de points amorces et le nombre de classes présumées sont les facteurs significatifs. La longueur de la multipartition donne un effet, quant à elle, qui se situe presque à la valeur critique, mais la nature des amorces n'est pas du tout significative, bien que présentant une légère supériorité. Ce qui infirme ce que notre première expérience laissait croire. Mais dans ce premier cas, on n'avait affaire qu'à deux classes, et l'intuition géométrique montre bien la supériorité des noyaux alignés dans ce cas. Avec la présence de plus de deux groupes (ici: 3 et 5), cet avantage disparaît. En fait on peut penser que les

	Facteurs				
Essais	1	2	3	4	5
1	-	-	-	-	+
2	-	-	-	+	-
3	-	-	+	-	-
4	-	-	+	+	+
5	-	+	-	-	-
6	-	+	-	+	+
7	-	+	+	-	+
8	-	+	+	+	-
9	+	-	-	-	-
10	+	-	-	+	+
11	+	-	+	-	+
12	+	-	+	+	-
13	+	+	-	-	+
14	+	+	-	+	-
15	+	+	+	-	-
16	+	+	+	+	+

Tableau 3.7: Le schéma de la deuxième expérience.

Essais	y_1	y_2	y_3	$\overline{y}$	s^2
1	.10	.18	.03	.103	.0056
2	.99	1	1	.996	.00003
3	1	1	.99	.996	.00003
4	.34	.32	.39	.35	.0013
5	0	0	0	0	0
6	.01	.01	.02	.013	.00003
7	.65	.62	.67	.635	.00045
8	.06	.05	.08	.063	.00023
9	1	1	1	1	0
10	0	0	0	0	0
11	.34	.39	.45	.393	.003
12	1	1	1	1	0
13	0	0	0	0	0
14	0	0	0	0	0
15	.29	.40	.46	.383	.086
16	.90	.74	.97	.87	.0139

Tableau 3.8: Résultats de la seconde expérience.

SOURCE	SS	DF	MS	F
Inter	7.933	15	.529	766.7
Intra	.221	32	.0069	
Totale	8.154	47		

Tableau 3.9: Analyse de la variance pour la seconde expérience.

Facteurs & Interactions	$\hat{I}$	$\hat{C}$	SS($\hat{C}$)	f($\hat{C}$)	% variation
1	.06125	.49	.0450	6.5285	.55
2	-.35925	-2.874	1.5497	224.25	19
3	.32225	2.578	1.2469	180.432	15.3
4	-.02725	-.218	.0089	1.287	.1
5	-.25925	-2.074	.8070	116.77	9.9
12	.07425	.594	.06619	9.578	.8
13	.08925	.714	.0956	13.8338	1.2
14	.05075	.406	.03092	4.474	.4
15	-.02075	-.166	.00517	0.75	.06
23	.20185	1.298	.316	45.8	3.9
24	.00925	.074	.00103	0.15	.01
25	.52725	4.218	3.3379	483.01	40.9
34	-.00375	-.03	.00017	0.02	.002
35	.21075	1.686	.5333	77.17	6.5
45	.05275	.422	.033	4.77	.4

Tableau 3.10: Calculs de signification des facteurs et pourcentages de variation expliquée. Deuxième expérience.

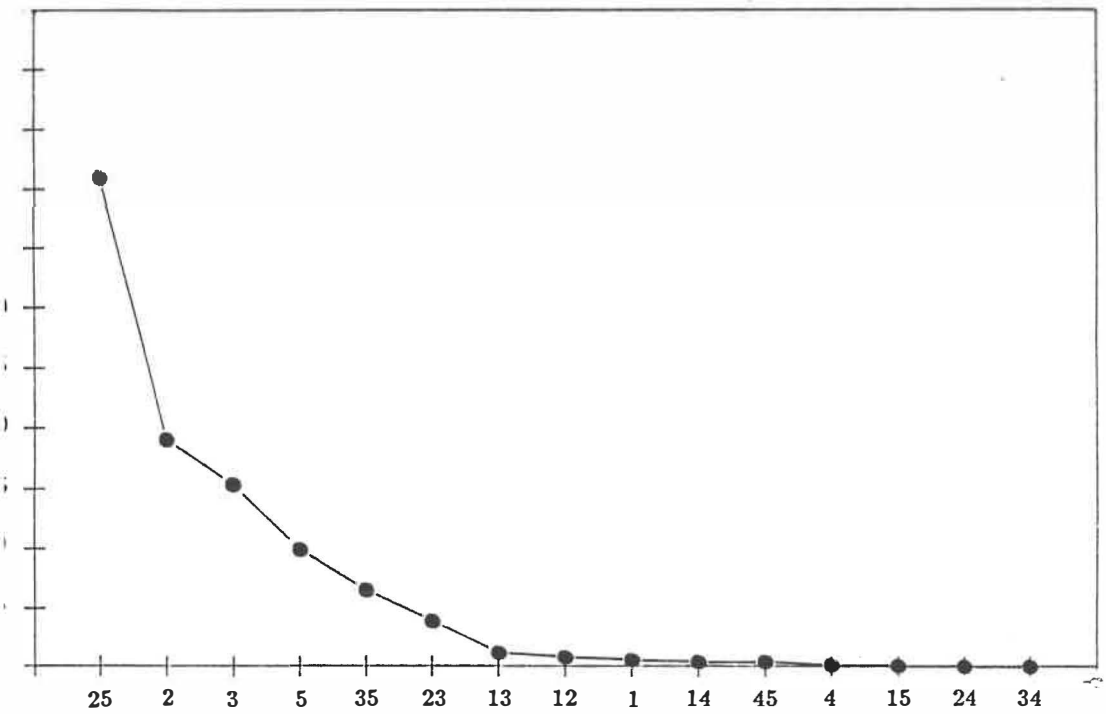

Figure 3.3: Diagramme de Pareto pour la deuxième expérience.

noyaux alignés pourraient être utilisés avec profit dans le cas où on a une absence de structure.

Les interactions avec la longueur de la multipartition qui sont significatives montrent l'avantage à utiliser une longueur 15, donc grande, sur une courte multipartition.

Une grande distance entre les groupes, facteur 3, favorise beaucoup la qualité du résultat. Il fallait s'y attendre!

L'interaction la plus intéressante, 25, la plus importante de tous les effets, montre ce qu'on sait déjà, à savoir qu'il y a intérêt a avoir une information *a priori* sur le nombre de classes en présence avant d'utiliser l'heuristique. En fait une sous-estimation du nombre de classes en présence, *i.e.* prendre moins de noyaux qu'il y a de groupes, donne un moins bon résultat qu'une légère surestimation.

Afin de donner plus de contenu à cette affirmation, nous avons isolé les deux facteurs: nombre de groupes en présence, et nombre de noyaux initiaux, et fait une petite expérience répétée trois fois dans les meilleures conditions de l'expérience précédente. Nous en trouvons les résultats au tableau 3.11.

Nous confirmons bien avec ces essais (l'analyse de variance n'est pas nécessaire), ce qu'on a déjà pressenti en ce qui concerne un petit nombre de groupes: il vaut mieux surestimer le nombre de groupes mais pas trop, que de le sous-estimer.

Mais en présence de sept groupes, les performances sont très mauvaises. On a exploré alors le facteur de densité, puisque, en présence de nombreux groupes dans le même espace, chacun contenant cinquante sujets, la densité est plus grande

# de noyaux	# de groupes				
	2	3	5	7	7
3	.90	1	.38	0.0	0.0
5	.83	.88	1	.07	.19
7	.56	.39	.96	.49	.99

Tableau 3.11: Moyenne sur trois essais pour l'interaction nombre de groupes et nombre de noyaux. La dernière colonne est obtenue avec des noyaux choisis au hasard.

qu'avec trois groupes par exemple. Ce qui laissait croire à plus de confusions. Mais des essais avec des groupes de vingt sujets n'ont pas amélioré les résultats de la colonne sept du tableau 3.11. On a noté ici que les noyaux alignés diminuaient beaucoup les performances en présence de nombreux groupes. On conçoit aisément que la contrainte des noyaux rectilignes produit en ceux-ci une mauvaise représentation des données. La dernière colonne reprend la colonne précédente avec, cette fois, les noyaux choisis au hasard. Les performances sont alors bien meilleures. On notera toutefois que pour des groupes aussi séparées, on aurait pu attendre bien mieux.

Ainsi donc, on ne peut pas aisément utiliser cette heuristique sans information sur ce qu'on attend... Le problème de la détermination du *bon* nombre de groupes reste entier, on ne voit pas bien comment acquérir plus de certitudes à ce sujet avec l'heuristique de Celeux.

Chapitre 4

Discussion

L'heuristique de Celeux fournit une méthode très rapide qui se prête fort bien à l'exploration des grands ensembles bien typés avec assez peu de groupes. Nous allons passer en revue un certain nombre des sujets abordés dans ce travail. Mais il faut d'abord renoter que nos données sont en général fort bien séparées et qu'on attendait un pourcentage de succès de l'heuristique plus élevé que celui qu'on a eu.

4.1 Les stratégies de validation

Notre choix d'utiliser des structures simples était guidé par un désir d'idéalisation de la réalité, afin d'extraire l'essence même des propriétés de performance de l'heuristique. D'être plus près de la théorie en somme. Nous n'avons examiné ici qu'un seul critère. Celui-ci est fort strict: au sens où la structure de données

devait être retrouvée intégralement pour donner à un essai la valeur 1. Ainsi une erreur sur un pourcentage infime de sujets entraîne la valeur 0. Dans les cas où les données deviennent plus denses dans leur espace de définition, cela peut jouer un rôle que nous avons négligé. On peut penser à d'autres critères de nature moins catégorique, examiner leurs propriétés.

4.2 Choix des facteurs

On a toujours travaillé en dimension deux, et même là la situation se complique rapidement. Lorsque le nombre de groupes augmente la densité des points dans l'espace de définition augmente. Cependant même avec 7 groupes la nature gaussienne considérée donne des groupes très séparés. Il y a encore une distance de 6.94 unités entre deux groupes adjacents des données heptagonales, et plus de 99.9% des points d'une gaussienne standard sont à distance inférieure à trois de la moyenne. Nous n'avons jamais considéré plus de cinquante points par groupe... De plus tous nos groupes ont même cardinalité.

Le facteur distributionnel n'a pas été exploré bien loin non plus: que des gaussiennes standard ou des uniformes sur le disque. Ce dernier choix imposé un peu pour des raisons d'étude de comportement des points alignés, dont nous reparlerons brièvement plus loin.

Pour une expérience de validation plus complète, on pourra consulter le diagramme d'Ishikawa(figure 4.1) illustrant un certain nombre de facteurs qui, on le pense, ont de l'effet sur la performance de l'heuristique de Celeux.

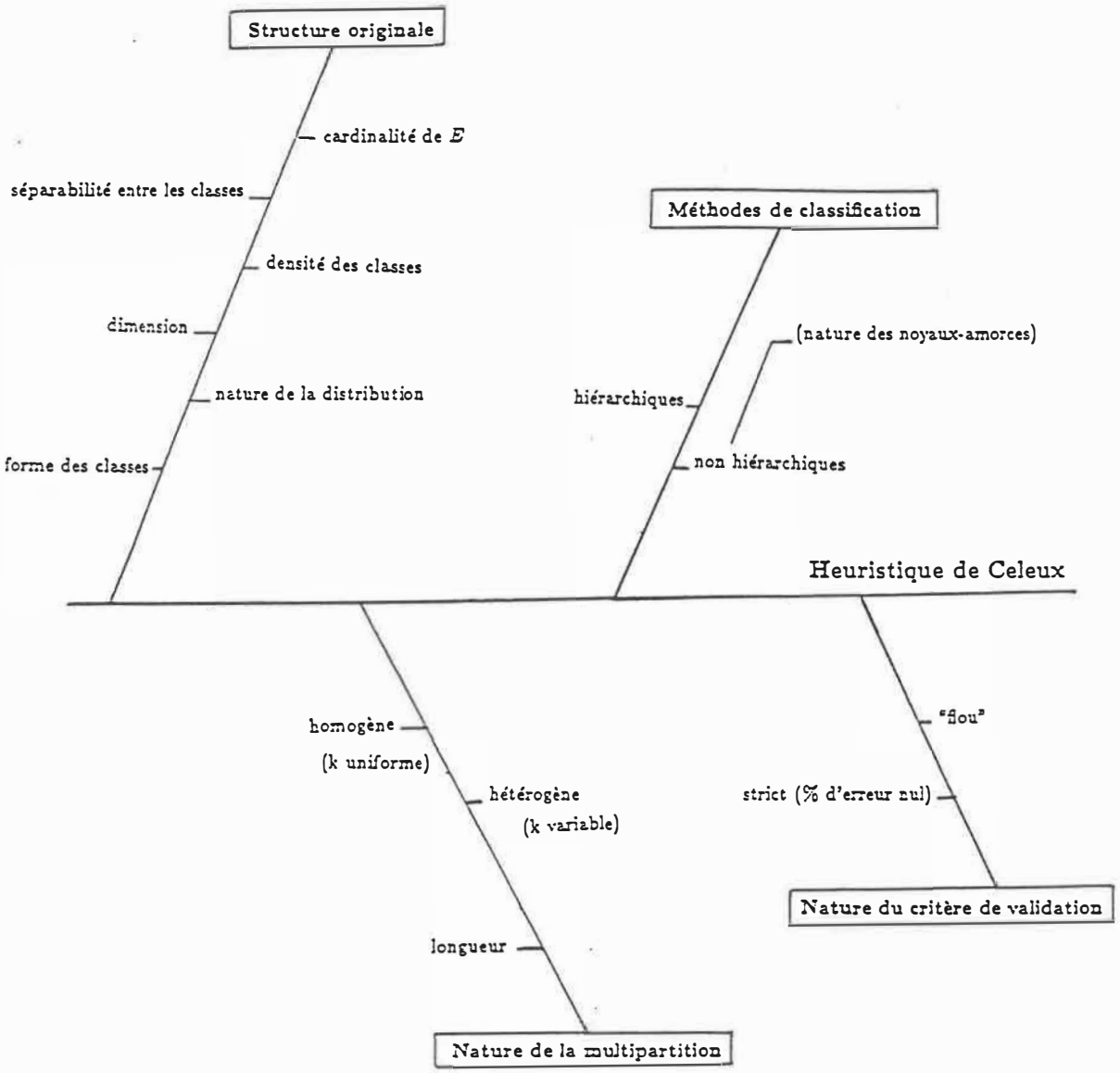

Figure 4.1: Diagramme d'Ishikawa pour l'heuristique de Celeux.

4.3 Interaction entre le nombre de groupes et le nombre d'amorces

Un des résultats escomptés et retrouvé avec beaucoup de force est cette interaction entre, en quelque sorte, le nombre réel de groupes et son nombre escompté qui sert de paramètre à l'heuristique. Pour que l'heuristique fonctionne bien, il faut avoir une bonne idée du véritable nombre de groupes présent dans les données. De plus, il y aurait lieu, en toute rigueur, de faire une exploration des concordances entre les partitions obtenues et la réalité (voir également la première section de cette discussion).

4.4 L'absence de structure

Une seule gaussienne standard n'éveille *a priori* aucun soupçon quant à la présence de plusieurs groupes. Or l'heuristique de Celeux est complètement impuissante, semble-t-il, à la retrouver. La question de la partition grossière comme partition centrale a fait l'objet ici d'une conjecture qui rendrait impossible l'utilisation des multipartitions et de leurs partitions centrales afin de détecter une telle absence de structure. On peut alors penser utiliser des caractéristiques distributionnelles de $C(u)$, liées par exemple au nombre k de groupes présumé, pour aller plus loin dans la direction du problème du "bon" nombre de groupes.

4.5 Les points alignés

Supposons qu'on construise une multipartition à partir de l'algorithme des centres mobiles sur une absence de structure, disons, pour simplifier, une loi uniforme sur le disque unité.

Il est facile de voir que des noyaux de départ alignés vont découper le disque en tranches parallèles. Que dire alors des distances entre paires de telles partitions? La question se pose: en général, quelles sont les formes stables de l'algorithme des centres mobiles? Peut-on exploiter ces propiétés pour détecter une absence de structure?

En ce qui nous concerne, on a vu que, sur les formes simples, là où il y a peu de groupes, des points alignés donnent des résultats légèrement meilleurs à l'heuristique de Celeux, que des noyaux randomisés. Mais avec sept groupes les noyaux alignés représentent mal la structure et donnent de mauvais résultats.

Bibliographie

[1] M.R. Anderberg. *Cluster Analysis for Applications.* Academic Press, New York, 1973.

[2] J. P. Barthélemy, B. Leclerc, and B. Monjardet. On the use of ordered sets in problems of comparison and consensus of classifications. *Journal of classification*, 3 no2:187–224, 1986.

[3] J.P. Barthélemy, B. Leclerc, and B. Monjardet. Quelques aspects du concensus en classification. In E. diday *et al.*, editor, *Data Analysis and Informatics III*, pages 307–316, North Holland, 1984.

[4] C.L. Bayne, Beauchamp John J., Begovich Connie L., and V.E. Kane. Monte carlo comparisons of selected clustering procedures. *Pattern Recognition*, 12:51–62, 1980.

[5] C.L. Begovich and V.E. Kane. Estimating the number of groups and group membership using simulation cluster analysis. *Pattern Recognition*, 15, n°4:335–342, 1982.

[6] G.E.P. Box, W.G. Hunter, and J.S. Hunter. *Statistics for Experimenters. An Introduction to Design, data Analsysi and model building.* John Wiley and Sons, New York, 1978.

[7] G. Celeux. *Approximation rapide et interprétation d'une partition centrale pour les algorithmes de partitionnement.* Technical Report 30, INRIA, Rocquencourt, 1984.

[8] G. Celeux, E. Diday, G. Govært, Y. Lechevallier, and H. Ralambondrainy. *Classification automatique des données.* Dunod, Paris, 1989.

[9] J.L. Chandon and S. Pinson. *Analyse typologique, théories et applications.* Masson, Paris, 1983.

[10] E. Diday. Le formalisme de base. In E. Diday et coll., editor, *Optimisation en classification automatique*, pages 10–25, INRIA, Rocquencourt, 1979.

[11] E. Diday, J. Lemaire, J. Pouget, and F. Testu. *Éléments d'analyse de données.* Dunod, Paris, 1982.

[12] R. Dubes and A. K. Jain. Validity studies in clustering methodologies. *Pattern Recognition*, 11:235–254, 1979.

[13] B.S. Everitt. *Cluster Analysis.* Halstead Heinemann, London, 1980.

[14] G. Govært. Algorithme de classification d'un tableau de contingence. 1980. Document interne INRIA, Rocquencourt.

[15] G. Govært. Classification croisée. 1983. Thèse d'État, Université Paris VI.

[16] A.K. Jain and J.V. Moreau. Bootstrap techniques in cluster analysis. *Pattern Recognition*, 20:547–568, 1987.

[17] B. Leclerc and G. Cucumel. Consensus en classifications: une revue bibliographique. *Mathématiques et Sciences Humaines*, 25:109–128, 1987.

[18] I.C. Lerman. *Les bases de la classification automatique.* Gauthier-Villars, collection programmation, Paris, 1970.

[19] F. Marcotorchino and P. Michaud. Agrégation de similarités en classification automatique. *Revue de Statistique Appliquée*, XXX, n°2:21–44, 1982.

[20] G.W. Milligan. An algorithm for generating artificial test clusters. *Psychometrika*, 50, n°1:123–127, 1985.

[21] G.W. Milligan and M.C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 5:159–179, 1985.

[22] Annie Morin. Comparaison de plusieurs méthodes de classification sur un exemple de lexicométrie. *Revue de Statistique Appliquée*, XXXII, n°4:37–49, 1984.

[23] Pascale Rousseau and Sankoff D. Analyse typologique de données binomiales; application en linguistique. In E. Diday et coll., editor, *Optimisation en classification automatique*, pages 429–442, INRIA, Rocquencourt, 1979.

[24] H. Späth. *Cluster Analysis algorithms for data reduction and classification of objects.* John Wiley and Sons, New York, 1980.

[25] M.G. Vigier. *Pratique des plans d'expériences, méthodologie de Taguchi.* Les éditions d'organisation, Paris, 1988.

[26] D.J. Wheeler. *Understanding Industrial experimentation.* Statistical Process Controls, Inc., 7026 Shadyland Drive, Knoxville, Tennessee 37919, 1988.

Annexe

Dans les pages qui suivent, nous donnons le programme écrit en Fortran77 qui a servi pour les analyses de ce mémoire.

Les données ont été obtenues à l'aide des générateurs de nombres aléatoires de la librairie IMSL.

PROGRAMME POUR CALCULER LA PARTITION

PSEUDO-CENTRALE PAR LE CRITERE DE CELEUX

Ce programme de nature interactif est conçu pour calculer la partition pseudo-centrale par l'heuristique de Celeux en utilisant l'algorithme de partionnement des centres mobiles. L'utilisateur est tenu à répondre à une serie de questions permettant de définir les paramètres suivants:
-le nombre d'observations:nobs
-le nombre de variables:nvar
Ces deux paramètres sont propres aux données qui sont lues à partir du fichier d'unité 10.
-la longueur de la multipartition:long
-le nombre de classes: k
C'est le nombre de points nécessaires pour amorcer l'algorithme des centres mobiles.
-la nature des points amorces:
On trouve deux options : alignés ou pris au hasard parmi les données.

```
/SYS REG=3000
/FILE SUBLIB PDS(*MUS,*OS,*MLIB,*SLIB)
/FILE FT05F001 TERM
/FILE FT10F001 NAME(KC02:N141)
/FILE FT13F001 NAME(KC02:RESLP) NEW(REPL) LRECL(132)
/FILE FT14F001 NAME(KC02:CCP) NEW(REPL) LRECL(132)
/FILE FT15F001 NAME(KC02:INDCP) OLD
/LOAD VSFORT


      INTEGER       NOBS,NVAR,K,LONG,U,V,W,Y,F,P,P1,NR,CENT(10)
      INTEGER        REP,CHOIX,LONG1,H
      DIMENSION     X(100,3),P(100),E(12),Q(12),S(12,3)
      DIMENSION     Y(100,20),F(100,100),X1(110,3),P1(110)
      REAL          DA,DB,DC

C     X   : LA MATRICE DES DONNEES
C     CENT:LE VECTEUR DE POINTS AMORCES POUR HMEANS
C     LONG:LA LONGUEUR DE LA MULTIPARTITION.
C     P   :P(I)=J  SIGNIFIE QUE L'INDIVIDU I
C           EST DANS LA CLASSE J .
C
C-------------LECTURE DES PARAMETRES-------------------
      WRITE(6,14)
14    FORMAT('QUEL EST LE NOMBRE D"OBSERVATIONS ?')
       READ(5,*) NOBS
      WRITE(6,16)
16    FORMAT('QUEL EST LE NOMBRE DE VARIABLES?')
      READ(5,*) NVAR
      WRITE(6,17)
17    FORMAT('QUEL EST LE NOMBRE DE CLASSES?')
```

```
      READ(5,*) K
      WRITE(6,18)
18    FORMAT ('QUEL EST LE NOMBRE DE PARTITIONS ?')
      READ(5,*) LONG
      WRITE(6,19)
19    FORMAT ('CHOISIR:',/,' 1. SI LES POINTS AMORCES SONT
     &        ALIGNES',/,
     &          '2. SI LES POINTS AMORCES SONT PRIS AU HASARD.')

      READ(5,*) CHOIX
C
      DO 10 IL=1,LONG
C
         IF (CHOIX.EQ.1) THEN
            CALL CENTALIG(X,NOBS,NVAR,K,NR,CENT,X1)
            CALL CONFIG (X1,NOBS,NR,NVAR,P1,P,K,CENT)
         ELSE
            CALL CENTHAS(K,NOBS,CENT,NVAR,X)
            CALL CONFIG (X,NOBS,0,NVAR,P,P,K,CENT)
         ENDIF


         CALL HMEANS(NOBS,NVAR,X,P,K,S,E,D,0)
C
         DO 20 I=1,NOBS,1
               Y(I,IL)=P(I)
20    CONTINUE
10    CONTINUE
C

         CALL FMF(LONG,NOBS,Y,F)
C
         CALL CENTRALE(NOBS,LONG,F,Y)
C
       STOP
       END
C

  C-------------------------------------------------------------|
  C  sous-programme HMEANS.                                     |
  C Ce module lit le vecteur P qui illustre la partition       |
  C initiale  et dans lequel retourne la partition optimale en |
  C utilisant  l'algorithme des centres mobiles.(SPATH ,1980)  |
  C                                                            |
  C-------------------------------------------------------------|


      SUBROUTINE HMEANS (NOBS,NVAR,X,P,K,S,E,D,IDR)

      DIMENSION     X(100,3),S(12,3),E(12),P(100),Q(12)
      INTEGER       IDR,NVAR,NOBS,K,R,Q,P,K1
      REAL          X,S,E,D
C
      ID=0
      DMAX=1.E30
```

```
      DO 1 I=1,NOBS
      R=P(I)
      IF(R.LT.1.OR.R.GT.K) RETURN
 1      CONTINUE
 2    DO 4     J=1,K
               Q(J)=0
               DO 3 K1=1,NVAR
                    S(J,K1)=0
 3      CONTINUE
 4      CONTINUE
        DO 6 I=1,NOBS
             R=P(I)
             Q(R)=Q(R)+1
             DO 5 K1=1,NVAR
                  S(R,K1)=S(R,K1)+X(I,K1)
 5       CONTINUE
 6       CONTINUE
         IR=0
         DO 8 J=1,K
              E(J)=0
              R=Q(J)
              F=0.
              IF(R.NE.0) F=1./FLOAT(R)
              IF (R.EQ.0)  IR=IR+1
              DO 7 K1=1,NVAR
              S(J,K1)=S(J,K1)*F
 7       CONTINUE
 8       CONTINUE
         D=0.
         DO 10 I=1,NOBS
               R=P(I)
               F=0.
               DO 9 K1=1,NVAR
                     T=S(R,K1)-X(I,K1)
                     F=F+T*T
 9       CONTINUE
         E(R)=E(R)+F
         D=D+F
 10      CONTINUE
         IF(IR.NE.0)  RETURN
         IF (D.GE.DMAX)    ID=ID+1
         IF (IDR.EQ.1)     WRITE(6,14) ID,D
         IF(ID.GT.15)  RETURN
         DMAX=0
         DO 13 I=1,NOBS
               F=1.E30
               DO 12 J=1,K
                     G=0.
                     DO 11 K1=1,NVAR
                            T=S(J,K1)-X(I,K1)
                            G=G+T*T
 11                  CONTINUE
                     IF(G.GE.F) GOTO 12
                     F=G
                     R=J
```

```
12              CONTINUE
                P(I)=R
13          CONTINUE
14          FORMAT(1X,'ID=',I4,4X,'D=',F18.8)
            GOTO 2
            END
```

```
C-------------------------------------------------------------|
C SOUS-PROGRAMME FMF.                                         |
C A partir de la matrice Y, dont les colonnes sont les vec -  |
C teurs Pl (1<=l<=long ),on calcule la matrice F des formes   |
C fortes qui n'est autre que le graphe de l'intersection des  |
C Pl.                                                         |
C     F(i,j)=1 ssi i et j sont classés ensemble dans Pl pour  |
C                tout l.                                      |
C     F(i,j)=0 sinon                                          |
C-------------------------------------------------------------|

      SUBROUTINE FMF(LONG,NOBS,Y,F)

      INTEGER   LONG,NOBS,Y,F
      DIMENSION  Y(100,20),F(100,100)

        DO 200 I=1,NOBS,1
               DO 300 J=1,NOBS,1
                      F(I,J)=0
                      DO 400 L=1,LONG,1
                             IF (Y(I,L).EQ.Y(J,L)) THEN
                                 F(I,J)=F(I,J)+1
                             ENDIF
400                    CONTINUE
                      IF (F(I,J).EQ.LONG) THEN
                         F(I,J)=1
                      ELSE
                         F(I,J)=0
                      ENDIF
300            CONTINUE
200       CONTINUE
          RETURN
          END
```

```
C-------------------------------------------------------------|
C sous-programme centrale.                                    |
C.A partir de la matrice Y sont calculées les formes fortes et|
C décrites sur le fichier d'unité 13.                         |
C.On calcule ,ensuite, la matrice B décrivant le nombre de fois|
C où les representants des formes fortes se retrouvent le long|
C de la multipartition.L'ensemble des representants est noté F.|
C.A partir de la matrice B,sont formées les m+1 relations U_l |
C définies par:                                               |
C      U_l(i,j)=1   ssi   B(i,j) > l                          |
C               =0   sinon                                    |
C.Recherche des fermetures transitives des U_l.               |
```

```
C.Calcul des distances entre les partitions ( l'heuristique de |
C  Celeux )                                                     |
C.Illustration de la partition obtenue.                         |
C---------------------------------------------------------------|
C===============================================================
           SUBROUTINE CENTRALE(NOBS,LONG,F,Y)
C---------------------------------------------------------------
       INTEGER NOBS,LONG,MAX,NFOUND1,I,J,MAX2,L,R
       INTEGER F(100,100),E(100),A1(100,100),FF(100),Y(100,20)
       INTEGER CARDF(100),B(100,100),U(20,100,100),E2(100)
       INTEGER W(20,100,100),CC(100),TRANS
       INTEGER   MIN ,IND ,CARDFC(100),FF2(100),A2(100,100)

        DO 115  I=1,NOBS,1
               E(I)=I
115    CONTINUE
        MAX=0
         DO 200 I=1,NOBS,1
              CARDF(I)=0
              IF (E(I).NE.0)  THEN
                MAX=MAX+1
                FF(MAX)=I
                DO 300 J=1,NOBS ,1
                       IF ((F(I,J).EQ.1).AND.(E(J).NE.0)) THEN
                          CARDF(MAX)=CARDF(MAX)+1
                          A1(MAX,CARDF(MAX))=J
                          E(J)=0
                       ENDIF
300             CONTINUE
      WRITE (13,30) FF(MAX),CARDF(MAX)
      WRITE (13,31) (A1(MAX,K),K=1,CARDF(MAX))
30    FORMAT(/,' FORM.FORT.REPRES.PAR':,I3,2X,/,'CARD:')
31    FORMAT(/,100I3)
             ENDIF
200   CONTINUE
      WRITE (13,20) MAX,(FF(I),I=1,MAX)
      WRITE (13,21) (CARDF(I),I=1,MAX)
20    FORMAT (//,'    LE NBR DE FORM.FORTES:',I5,/,100I3)
21    FORMAT(100I3)

C---------------------------------------------------------------
C      CONSTRUCTION DE LA MATRICE B DES FORMES FORTES.
C
C      W(K) MATRICE ASSOCIEE A LA PARTITION PK
C---------------------------------------------------------------
       DO 990 L=1,LONG,1
            DO 990 I=1,MAX
              DO 990 J=1,MAX
               W(L,I,J)=0
990    CONTINUE
       NFOUND1=0
       DO 220 I=1,MAX,1
          DO 230 J=1,MAX,1
            DO 240 L=1,LONG,1
               IF (Y(FF(I),L).EQ.Y(FF(J),L))  THEN
```

```
                NFOUND1= NFOUND1 + 1
                W(L,I,J)=1
              ELSE
                W(L,I,J)=0
              ENDIF
240     CONTINUE
          B(I,J)=NFOUND1
          NFOUND1 = 0
230     CONTINUE
        WRITE(14,81) (B(I,J),J=1,MAX)
220     CONTINUE

C-----------------------------------------------------------------
C            RECHERCHE DES  RELATIONS U1 SUR F
C         U1(F,F')=1       SI      B(F,F')>=1
C         UJ(F,F')=0 SINON
C-----------------------------------------------------------------

        DO 999 K=1,LONG+1,1
            DO 999 I=1,MAX,1
              DO 999  J=1,MAX,1
                U(K,I,J)=0
999           CONTINUE
        DO 400  K=1,LONG+1,1
               KK=K-1
               WRITE(14,80) K
80      FORMAT (/,'K=',I3)
        DO 400  I=1,MAX,1
              DO 420  J=1,MAX,1
                IF (B(I,J).GE.KK) THEN
                  U(K,I,J)=1
                ENDIF
420            CONTINUE
        WRITE (14,81) (U(K,I,J),J=1,MAX)
400     CONTINUE

C-----------------------------------------------------------------
C             RECHERCHE DES FERMETURES TRANSITIVES
C                DES  U1 CONTRUITES SUR F.
C-----------------------------------------------------------------

        WRITE (14,82)
82      FORMAT(//,'LES FERMETURES TRANST. ASSOCIEES AUX U(K):')
        DO 520 K=1,LONG+1,1
            DO 520 TRANS=1,2
               DO 520 I=1,MAX,1
                  DO 520 J=1,MAX,1
                     DO 510 L=1,MAX,1
C
                IF ((U(K,I,J).EQ.1).AND.(U(K,J,L).EQ.1)) THEN
                    U(K,I,L)=1
                    U(K,L,I)=1
                ENDIF
510             CONTINUE
520    CONTINUE
```

```
81     FORMAT(100I3)
C-------------------------------------------------------------------
C          CALCUL DES DISTANCES ENTRE LES RELATIONS
C-------------------------------------------------------------------
       DO 600 K=1,LONG+1,1
            WRITE(14,80) K
            DO 601 I=1,MAX,1
                WRITE(14,81) (U(K,I,J),J=1,MAX)
601         CONTINUE
            CC(K)=0
            DO 620 L=1,LONG,1
                DO 620 I=1,MAX,1
                    DO 620 J=1,MAX,1
                      R=CARDF(I)*CARDF(J)
                      CC(K)=CC(K)+R*ABS(U(K,I,J)-W(L,I,J))
620                 CONTINUE
            CC(K)=0.5*CC(K)
            WRITE (14,619) K,CC(K)
 619        FORMAT(' K=',I3,5X,I10)
 600   CONTINUE

       IND=1
       MIN = CC(1)
       DO 602 K=1,LONG+1,1
             IF (CC(K).LT.MIN) THEN
                 MIN=CC(K)
                 IND=K
             ENDIF
602    CONTINUE
            WRITE(14,*)'IND=',IND
            WRITE (14,*) 'MIN=',MIN

C==========LE NOMBRE DE CLASSES INDIQUEES PAR LA============
C========== PARTITION RETENUE PAR L'HEURISTIQUE=============

       DO 715  I=1,MAX,1
                 E2(I)=I
715    CONTINUE

       MAX2=0
       DO 800 I=1,MAX,1
           CARDFC(I)=0
           IF (E2(I).NE.0)  THEN

               MAX2=MAX2+1
               FF2(MAX2)=I
                  DO 720 J=1,MAX ,1
                  IF ((U(IND,I,J).EQ.1).AND.(E2(J).NE.0)) THEN
                     CARDFC(MAX2)=CARDFC(MAX2)+1
                     A2(MAX2,CARDFC(MAX2))=J
                     E2(J)=0
                  ENDIF
720               CONTINUE

      WRITE(14,251)FF2(MAX2),CARDFC(MAX2)
```

```
      WRITE(14,252) A2(MAX2,K),K=1,CARDFC(MAX2))
251   FORMAT('CLASSE DES FORMES FORTES:',I3,2X,/,'CARD:',I3)
252   FORMAT(/,100I3)

          ENDIF

 800  CONTINUE

      WRITE(15,*)  MAX2  ,MIN
      WRITE(14,16) MAX2,(FF2(I),I=1,MAX2)
      WRITE(14,17) (CARDFC(I),I=1,MAX2)
16    FORMAT('  LE NB DE CLAS:',I5,/,' CLAS:',100I5)
17    FORMAT(' CARD:',100I5)

      DO 252 I=1,MAX2
          WRITE(14,18)I
18        FORMAT('  CLASSE:',I5)

          DO 253 J=1,CARDFC(I)
             WRITE(14,19) (A1(A2(I,J),L),L=1,CARDF(A2(I,J)))
19           FORMAT (10I5)
253       CONTINUE

252       CONTINUE
      END

C-------------------------------------------------------------------
C Sous-programme centalig.
C Il fait appel à la librairie IMSL pour faire un choix au hasard
C de deux indices parmi nobs.Ceux-ci sont les premiers éléments
C du vecteur CENT ; k-2 points sont choisis au hasard dans le
C segment joignant les premiers.On leur attribue k-2 lignes dans
C la matrice X1 numérotées de nobs+1 jusqu'à nobs+k-2.Ces numéros
C constituent le reste des composantes du vecteur CENT.
C-------------------------------------------------------------------

C------------PROGRAM POUR FOURNIR LES POINTS AMORCES-------------
C-------------------------ALIGNES-------------------------------

        SUBROUTINE CENTALIG(X,NOBS,NVAR,K,NR,CENT,X1)

        INTEGER    I,IDO,J,LDX,LDSAMP
        INTEGER    K,NX,NSAMP,INDEX,NR,CENT
        REAL       SAMP(2,2),R(10),X(100,3),X1(110,3)
        DIMENSION INDEX(2),CENT(10)
        EXTERNAL   RNSRS,RNUN
C
      LDX=NOBS
      NSAMP=2
      LDSAMP=2
      NR=K-2
C
      OPEN(10)
      DO 10 I=1,NOBS
           READ (10,11) (X(I,J),J=1,NVAR)
```

```
11          FORMAT(1X,2F12.4)
10      CONTINUE
        IDO=0
        CALL RNSRS(IDO,NOBS,NVAR,X, NOBS,2,NX,SAMP,LDSAMP,INDEX)

        CENT(1)=INDEX(1)
        CENT(2)=INDEX(2)
        DO 110 I=1,NOBS
            DO 110 J=1,NVAR
                X1(I,J)=X(I,J)
110     CONTINUE

       IF ( NR.GT.0) THEN
          CALL RNUN(NR,R)
          DO 120 I=1, NR
             DO 120 J=1, NVAR
               CENT(I+2)=NOBS+I
         X1(I+NOBS,J)=R(I)*X(INDEX(1),J)+(1-R(I))*X(INDEX(2),J)
120       CONTINUE
       ENDIF

       CLOSE(10)
       RETURN
       END

C-------------------------------------------------------------
C Sous-programme centhas,
C on choisit k indices au hasard parmi nobs(IMSL) qui vont
C constituer les composantes du vecteur CENT.
C-------------------------------------------------------------

 C-------------CHOIX DE POINTS AMORCES AU HASARD--------------

        SUBROUTINE CENTHAS(K,NOBS,CENT,NVAR,X)
        INTEGER    CENT(10),NOBS,K,NVAR
        DIMENSION  X(100,3)
        EXTERNAL   RNSRI
 C
        CALL RNSRI(K,NOBS,CENT)
        OPEN(10)
        DO 12 I=1,NOBS
           READ(10,11)   (X(I,J),J=1,NVAR)
11         FORMAT(1X,2F12.4)
12      CONTINUE
        CLOSE(10)
        RETURN
        END

C-------------FORMATION DES CLASSES AUTOUR DES POINTS AMORCES---

C-------------------------------------------------------------
C Sous-programme CONFIG.
C Après la lecture des données et le choix du vecteur CENT soit
C CENTALIG soit par CENTHAS,le sous-programme est appelé pour
C obtenir une configuration initiale pour amorcer la procédure
```

```
C HMEANS .Ce programme forme une partition en k classes autour
C des points amorces.Celle-ci est illustrée dans le vecteur P.
C-----------------------------------------------------------------
C
      SUBROUTINE CONFIG (X1,NOBS,NR,NVAR,P1,P,K,CENT)
C
      INTEGER    NOBS,NVAR,K,P,P1,NR,CENT(10),L
      REAL       DA,DB,DC
      DIMENSION  X1(110,3),P(100),P1(110)
C
      DO 40 I=1,NOBS+NR
            DA=1.0E10
                DO 30 L=1,K
                  DB=0.0
                  DO 20 J=1,NVAR
                     DC=X1(I,J)-X1(CENT(L),J)
                     DB=DB+DC*DC
                     IF (DB.GE.DA) GOTO 30
20                CONTINUE
                DA=DB
                P1(I)=L
30              CONTINUE
40    CONTINUE
      DO 12 I=1,NOBS
          P(I)=P1(I)
12    CONTINUE
      RETURN
      END
```