5

CLASSIFICATION DES DONNEES D'ENQUETES

Maurice Roux

Université Marseille 3
Saint-Jérôme

5.1 Introduction

5.1.1 Objectif et définitions

Comme dans d'autres méthodes d'analyse des données, la classification traite des tableaux de données rectangulaires, où les lignes représentent généralement les individus et les colonnes des variables ou des questions. Le but de telles méthodes est de découvrir des structures cachées de l'ensemble des individus, ces structures étant des groupes ou des hiérarchies de groupes emboîtés.

Ceci suppose que de telles structures existent réellement, mais on donne rarement un sens très précis à cette notion. La plupart des méthodes commencent par construire un tableau de distances, ou dissimilarités, inter-individuelles. La formule pour calculer ces distances est parfois choisie par l'utilisateur, parfois elle fait partie de la méthode (même si cela n'est pas toujours explicite). Les résultats de la classification dépendent largement de cette formule.

Certaines méthodes ne visent qu'à établir une partition des individus en un certain nombre de groupes, ce nombre étant généralement choisi par l'utilisateur. On désire que les groupes formés soient aussi homogènes que possible, c'est à dire que leurs individus soient très ressemblants entre eux, tandis que deux individus appartenant à des groupes différents doivent être très dissemblants.

Les méthodes hiérarchiques cherchent à établir des groupes d'individus similaires puis des assemblages de groupes formant des "super-groupes" eux mêmes réunis dans des classes plus hétérogènes, etc.

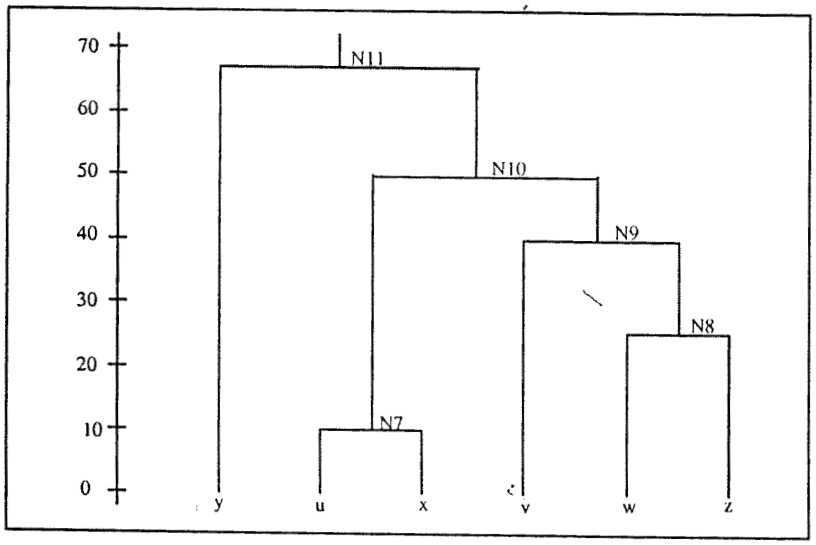

Figure 1

Un arbre hiérarchique ou dendrogramme
Les nœuds de l'arbre sont numérotés à partir de n + 1 lorsque n est le nombre d'individus (ici n = 6) : N7, N8, N9, N10, N11.
Le nœud N11 est aussi appelé "racine" ou "sommet" de l'arbre.

Un arbre hiérarchique, ou dendrogramme, est habituellement dessiné pour représenter ces emboîtements de groupes, comme dans la figure 1, où les individus sont placés à l'extrémité inférieure des branches de l'arbre. L'échelle verticale sur la gauche de ce dendrogramme représente une mesure de l'éloignement des groupes les uns par rapport aux autres. Ainsi la distance entre le groupe {u,x} et le groupe {v,w,z} est 50. Cette distance doit être considérée comme une moyenne des distances entre individus de l'un des groupes et ceux d'un autre. Remarquons qu'à partir de cette figure on peut déterminer des distances entre deux individus quelconque en prenant la valeur du nœud le plus bas qui couvre les deux individus considérés.

Tableau 1

Distance ultramétrique associée au dendrogramme de la figure 1

	u	v	w	x	y	z
u	0					
v	50	0				
w	50	40	0			
x	10	50	50	0		
y	65	65	65	65	0	
z	50	40	25	50	65	0

Une telle distance (cf. Tableau 1) a la propriété suivante pour tout ensemble de trois individus :

$$d(x,z) \leq \text{Max} \{ d(x,y), d(y,z) \}$$

Cette expression, appelée inégalité ultramétrique, est équivalente à dire que tous les triangles sont isocèles avec la base inférieure aux côtés égaux. Une distance ayant cette propriété est dite distance ultramétrique, ou ultramétrique en abrégé. L'inégalité ci-dessus est plus contraignante que l'inégalité triangulaire classique :

$$d(x,z) \leq d(x,y) + d(y,z)$$

C'est à dire que toute ultramétrique satisfait à l'inégalité triangulaire. De plus on peut démontrer qu'il y a une correspondance biunivoque entre les ultramétriques et les dendrogrammes.

5.1.2 Problèmes de complexité

Concentrons-nous temporairement sur l'obtention de partitions, et plus particulièrement sur les partitions en deux classes seulement, que nous appellerons des bipartitions. On peut facilement mettre au point une mesure de l'accord entre une partition et un tableau de distances initiales données. Mais en l'état actuel de nos connaissances il n'existe pas de manière simple d'obtenir la meilleure bipartition possible de l'ensemble des individus.

Cette particularité tient au fait que l'ensemble des bipartitions est très grand et que l'on ne connaît pas de propriété mathématique qui permettrait d'éviter l'examen de toutes ces bipartitions. Si l'on a n individus alors le nombre de bipartitions est $2^{n-1} - 1$ qui croît exponentiellement avec n (Edward et Cavalli-Sforza, 1965). Le nombre de structures d'arbre hiérarchique (indépendantes de la valeur des niveaux de nœuds) est encore beaucoup plus grand.

De sorte que l'on est obligé de recourir à des procédures heuristiques qui produiront une partition ou une hiérarchie en bon accord avec les données, mais sans donner, en général, le meilleur résultat possible.

Cet état de fait a quelques conséquences fâcheuses : il existe un très grand nombre de méthodes, chacune ayant en outre un certain nombre d'options et de variantes. L'utilisateur a donc à faire un choix difficile. De plus ces méthodes sont souvent très sensibles à l'échantillonnage : la suppression ou l'adjonction d'un seul individu peut conduire à des résultats différents. Enfin il est difficile d'établir la validité d'une classification ; même si les données se présentent comme un continuum, les méthodes de classification

fourniront des classes parce qu'elles sont faites pour cela, mais elles ne diront pas grand chose sur la validité des classes obtenues.

Dans le paragraphe 5.2 nous examinerons les grandes catégories de procédures classificatoires. Ensuite (§ 5.3) nous étudierons les particularités liées au dépouillement d'enquêtes, puis la détermination du rôle des variables dans le regroupement des individus (§ 5.4). Enfin nous conclurons en rappelant les stratégies les mieux adaptées au traitement d'enquêtes.

5.2 Les principales méthodes usuelles en classification

5.2.1 Classification hiérarchique

La construction de dendrogrammes de façon automatique a débuté dans les années 1960 (Sokal et Sneath, 1963). Deux types de méthodes existent : soit on construit l'arbre à partir du "bas", c'est à dire en agglomérant les individus les plus proches, puis en réitérant le processus sur les groupes obtenus ; soit en commençant par le "haut", c'est à dire en procédant par subdivisions successives de l'ensemble à classer.

Dans le reste de ce paragraphe nous supposerons que tout le travail préliminaire sur les données, comme le codage, la standardisation, est fait et que l'on a calculé une matrice de distances inter-individuelles à partir de laquelle on va élaborer la hiérarchie. Il ne faut cependant pas dissimuler que ce travail préliminaire est probablement celui qui a le plus d'influence sur les résultats de la classification.

5.2.1.1 Méthodes agglomératives élémentaires

Dans les méthodes agglomératives on répète alternativement deux étapes principales. La première consiste à balayer la matrice des distances en recherchant la paire d'individus (ou de groupes) les plus proches. Cette paire forme alors un nouveau groupe (ou "super-groupe"). La deuxième étape est le recalcul des distances entre le groupe nouvellement formé et le reste des individus ou groupes, pour obtenir une nouvelle matrice de distances réduite d'une ligne et d'une colonne.

	a	b	c	d	e
a	0				
b	25	0			
c	18	30	0		
d	25	40	10	0	
e	10	34	15	18	0

	a	b	c-d	e
a	0			
b	25	0		
c-d	21.5	35	0	
e	10	34	16.5	0

Pas 2

	a-e	b	c-d
a-e	0		
b	29.5	0	
c-d	19	35	0

Pas 3

	a-e c-d	b
a-e c-d	0	
b	32.25	0

Pas 4

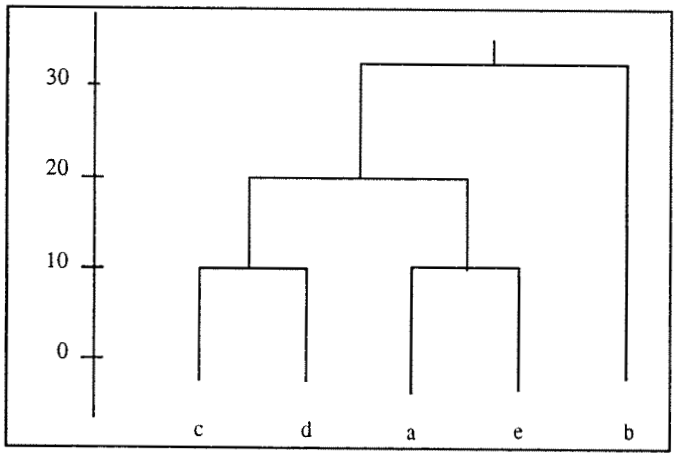

Figure 2

Détail des calculs pour la construction agglomérative d'une hiérarchie par la méthode du lien moyen

On réitère ces deux étapes alternativement jusqu'à ce que tous les groupes aient été fusionnés pour n'en former plus qu'un seul identique à l'ensemble de départ. L'agrégation de deux individus, ou groupes, est représentée par la jonction des deux branches correspondantes de l'arbre hiérarchique, la hauteur de cette jonction étant proportionnelle à la valeur de la distance entre les deux individus, ou groupes, en question.

La seule difficulté dans cette procédure réside dans le choix d'une formule raisonnable pour le recalcul des distances entre groupes. La formule la plus classique est certainement celle de la distance moyenne :

$$d(F, G_1 \cup G_2) = [n_1\, d(F, G_1) + n_2\, d(F, G_2)] / (n_1 + n_2) \qquad (1)$$

dans laquelle les lettres ont les significations suivantes :

- G_1 et G_2 sont les deux groupes que l'on fusionne,

- n_1 et n_2 sont les effectifs de ces deux groupes,
- d (F, G) est la distance entre les deux groupes F et G,
- $G_1 \cup G_2$ représente la réunion des deux groupes G_1 et G_2.

Le rôle des coefficients n_1 et n_2 est de prendre en compte le fait que les anciennes distances d (F, G_1) et d (F, G_2) avaient elles-mêmes été calculées sur n_1 et n_2 individus respectivement. Cette formule a été utilisée dans l'exemple de la figure 2. Dans les ouvrages de langue anglaise l'algorithme associé à cette formule est souvent appelé UPGMA, écriture abrégée pour "Unweighted Pair Group Method of Aggregation". Le terme "Unweighted" prête à confusion étant donnée la présence des pondérations n_1 et n_2 dans la formule. Il signifie en fait que, dans la nouvelle distance, toutes les distances inter-individuelles jouent le même rôle.

D'autres formules couramment utilisées sont celles du lien simple, ou du saut minimum (Sneath, 1957) :

$$d(F, G_1 \cup G_2) = \text{Min}, [\ d(F, G_1)\ ;\ d(F, G_2)\ ] \qquad (2)$$

et celle du lien complet, ou du diamètre (Sorensen, 1948) :

$$d(F, G_1 \cup G_2) = \text{Max}\ [\ d(F, G_1)\ ;\ d(F, G_2)\ ] \qquad (3)$$

Si l'on a en tête une représentation "factorielle" des données, on peut dire que la méthode du lien simple détecte bien des groupes étroits et allongés, tandis que celle du lien complet recherche plutôt des groupes compacts et sphériques. Cependant, en pratique, on ne connaît pas à l'avance la forme des groupes que contiennent les données. De plus, dans le cas où il n'y a pas de groupe très individualisé, la méthode du lien simple a tendance à agréger à la première paire réunie les autres individus, un par un, jusqu'au stade final où tous les individus sont réunis, empêchant ainsi la découverte de toute structure dans les données. Cet inconvénient est appelé "Effet de chaîne". En contrepartie l'algorithme du lien complet a tendance à couper en deux tout sous-ensemble ayant la moindre tendance à être allongé. Tout bien considéré la méthode du lien moyen constitue un bon compromis entre ces deux extrêmes.

5.2.1.2 La méthode du moment d'ordre 2, ou méthode de Ward

La méthode de Ward (Ward, 1963) est plus souvent appelée en français "méthode du moment d'ordre deux" (Benzécri et coll., 1973). Elle est basée sur la généralisation multidimensionnelle de l'équation usuelle de l'Analyse de la Variance. Cela suppose que les observations soient

(préalablement) réparties en un certain nombre de groupes s'excluant mutuellement de façon à former une partition P :

$$SCE_{tot} = SCE_{res} + SCE_{fac} \qquad (4)$$

- SCE_{tot} = Somme des carrés des écarts totaux,
- SCE_{res} = Somme des carrés des écarts résiduels,
- SCE_{fac} = Somme des carrés des écarts factoriels.

Dans cette formule les écarts sont relatifs à la moyenne des observations. Ainsi dans SCE_{tot} il s'agit de la moyenne générale, dans SCE_{res} il s'agit des écarts par rapport aux moyennes de chaque groupe, et dans SCE_{fac} on considère les écarts entre les moyennes des groupes et la moyenne générale. Ces sommes de carrés d'écarts s'appellent aussi inerties, par analogie avec les calculs effectués en Mécanique.

L'équation (4) ci-dessus signifie que l'inertie totale peut se décomposer en la somme des inerties intra-groupes et de l'inertie inter-groupe. D'une façon plus précise, si l'on appelle x la (seule) variable mesurée, z la moyenne générale de cette variable sur l'échantillon considéré, z_p la moyenne sur le groupe p, la formule (4) peut être réécrite :

$$\sum_{i \in I} (x_i - z)^2 = \sum_{p \in P} \sum_{i \in p} (x_i - z_p)^2 + \sum_{p \in P} n_p (z_p - z)^2 \qquad (5)$$

où n_p est le nombre d'individus dans le groupe p de la partition P. Mais les équations (4) ou (5) sont encore vraies dans le cas où l'on a plusieurs variables, à condition de remplacer les moyennes par les centres de gravité, et les sommes de carrés d'écarts par les sommes de carrés de distances entre observations et centres de gravité :

$$\sum_{i \in I} d^2 (x_i, z) = \sum_{p \in P} \sum_{i \in p} d^2 (x_i, z_p) + \sum_{p \in P} n_p d^2 (z_p, z) \qquad (6)$$

Dans cette équation les lettres z et z_p désignent respectivement le centre de gravité général et le centre de gravité du groupe p ; les distances, symbolisées par la lettre d, doivent être calculées selon une formule euclidienne. Naturellement, pour que ces notions aient un sens, il faut que les variables observées sur chaque individu soient de nature quantitative, mais nous verrons plus loin (§ 5.3.1) comment contourner cette difficulté avec les données d'enquêtes qui sont généralement qualitatives.

Si on appelle x_{ij} la valeur de la j-ème variable mesurée sur l'individu i, alors le centre de gravité général, z, est un vecteur dont les composantes sont :

$$z_j = (1/n) \sum_{i \in I} x_{ij} \qquad (7)$$

De la même façon les centres de gravité des groupes sont également des vecteurs ainsi définis :

$$z_{pj} = (1/n_p) \sum_{i \in p} x_{ij} \qquad (8)$$

Avec ces notations la distance euclidienne usuelle s'écrit :

$$d^2(x,y) = \sum_{j \in J} (x_j - y_j)^2 \qquad (9)$$

avec J désignant l'ensemble des variables mesurées.

S'il existe des groupes au sein de l'ensemble des individus étudiés, alors on devrait avoir une inertie intra-groupe faible donc une inertie inter-groupe forte, puisque la somme de ces deux quantités est l'inertie totale qui ne dépend pas d'une quelconque partition. Bien qu'on puisse aussi bien utiliser l'inertie intra-groupe, on prendra généralement l'inertie inter-groupe comme critère d'une bonne classification ; ou encore le rapport de l'inertie inter-groupe à l'inertie totale exprimé en pourcentage qui sera appelé le pourcentage d'inertie expliquée par la partition. Il sera largement fait appel à ces notions dans les méthodes non-hiérarchiques (§ 5.2.2), mais voyons d'abord comment elles peuvent être utilisées avec profit en classification hiérarchique.

La méthode du moment d'ordre deux, comme les méthodes étudiées précédemment, procède par agrégations successives. Examinons donc ce qui se passe pour l'inertie lorsque l'on fusionne deux classes p et p' à une étape quelconque de l'algorithme. Nous pouvons appliquer la formule (6) au sous-ensemble $p \cup p'$ et à sa partition en deux, définie par les deux groupes p et p' :

$$\sum_{i \in p \cup p'} d^2(x_i, z_{p \cup p'}) = \sum_{i \in p} d^2(x_i, z_p) + \sum_{i \in p'} d^2(x_i, z_{p'}) + n_p\, d^2(z_p, z) + n_{p'}\, d^2(z_{p'}, z)$$

Les deux sommations, à droite du signe égal, représentent les inerties intra-groupes, tandis que les deux derniers termes représentent l'inertie inter-groupe. Si les deux groupes p et p' sont agrégés, l'expression ci-dessus deviendra partie intégrante de l'inertie intra-groupe, en lieu et place des deux premiers termes qui contribuent à l'inertie intra-groupe avant la fusion. On voit donc que la nouvelle inertie intra-groupe sera supérieure à la somme des inerties intra-groupes avant agrégation, cette augmentation étant évaluée par :

$$D^2(p, p') = n_p\, d^2(z_p, z) + n_{p'}\, d^2(z_{p'}, z) \qquad (10)$$

Le principe de la méthode du moment d'ordre deux est de choisir, comme paire de groupes à agréger, celle qui minimise cette quantité.

Certains programmes utilisent le cadre général des méthodes agrégatives décrites ci-dessus, pour calculer la hiérarchie du moment d'ordre deux, en remplaçant les distances par l'accroissement de l'inertie intra-groupe, telle qu'elle est définie par (10). Le seul problème est alors de recalculer ces accroissements d'inerties après la fusion de deux groupes. Mais on peut démontrer qu'il est possible d'écrire les nouveaux accroissements d'inertie sous forme d'une expression ne contenant pas explicitement les centres de gravité z_p des groupes mis en jeu :

$$\begin{aligned} D^2(p \cup p', q) &= D^2(p, q)\,(n_{p'} + n_q) / (n_p + n_{p'} + n_q) \\ &+ D^2(p', q)\,(n_{p'} + n_q) / (n_p + n_{p'} + n_q) \\ &- D^2(p, p')\, n_q / (n_p + n_{p'} + n_q) \end{aligned}$$

Cependant cette façon de faire n'est pas efficace pour le traitement d'enquêtes. En effet, le point de départ est dans ce cas une matrice de distances inter-individuelles qui peut être énorme. Par exemple si l'enquête porte sur 1 000 individus, le tableau des distances comportera (1 000 x 999)/2 cases, soit 499 500 valeurs. Mais, si le nombre des variables est de 20, le tableau rectangulaire des données ne comportera que 20 000 cases, ce qui est donc beaucoup plus facile à gérer dans la mémoire de l'ordinateur. En contre-partie il faudra recalculer des distances entre individus, ou entre groupes, à chaque fois qu'on en aura besoin. On peut montrer que l'expression (10) peut se mettre sous la forme simplifiée suivante :

$$D^2(p, p') = \left[n_p\, n_{p'} / (n_p + n_{p'}) \right] d^2(z_p, z_{p'}) \qquad (11)$$

dans laquelle le centre de gravité des deux groupes, $z_{p \cup p'}$, a disparu. On peut donc travailler sur le tableau rectangulaire des individus par les variables, dans lequel on remplace progressivement, au fur et à mesure des agrégations, certains groupes de lignes par leur centre de gravité. Ainsi programmée cette méthode est particulièrement efficace pour le dépouillement d'enquêtes.

Il y a cependant quelques inconvénients à cette méthode. Tout d'abord le niveau des noeuds de la hiérarchie obtenue représente un accroissement d'inertie, proportionnel à un carré de distances. Cela produit une élongation excessive des branches vers le haut, et un "tassement" des niveaux vers le bas de l'arbre, ce qui fait apparaître les groupes de bas niveau comme beaucoup plus homogènes et séparés les uns des autres qu'ils ne sont en réalité. D'autre part il faut signaler la tendance de cette méthode à faire des groupes sphériques et bien équilibrés en effectifs, donc sa faible capacité à détecter des individus isolés, ou des groupes un peu étirés.

5.2.1.3 Les méthodes divisives

Les méthodes hiérarchiques divisives commencent par séparer l'ensemble complet en deux sous-ensembles, puis de nouveau chacun de ces sous-ensembles est partagé en deux classes, et ainsi de suite jusqu'à ce que l'on atteigne des classes ne contenant plus qu'un seul individu.

Pour travailler correctement on devrait logiquement faire le choix préalable d'un critère de bonne partition, par exemple le moment d'ordre deux interclasse, déjà examiné ci-dessus. Ensuite, pour chaque étape, c'est à dire pour chaque classe contenant plus d'un individu, on devrait examiner toutes les subdivisions de cette classe en deux sous-classes. Enfin il faudrait retenir la subdivision qui donne la valeur optimale du critère choisi. Cependant, comme on l'a expliqué au paragraphe 5.1.2, cette procédure est excessivement longue et devient même impraticable dès que l'on atteint une vingtaine d'individus.

Certains auteurs ont cherché à contourner cette difficulté. Ainsi Hubert (1973) a proposé de prendre pour noyaux des deux nouvelles classes les points les plus éloignés de cette classe, puis de répartir les individus restants, suivant diverses procédures progressives, entre les deux classes en fonction de leurs distances aux noyaux. Mais aucune des variantes possibles de cet algorithme n'a donné de résultats convaincants, probablement parce que les deux individus les plus éloignés d'une classe n'ont pas grand'chose à voir avec la structure globale de la classe.

Macnaughton-Smith et al. (1964) ont proposé un algorithme voisin, qui consiste à extraire de la classe à scinder le point dont la distance moyenne aux autres est la plus grande. On cherche ensuite à extraire d'autres individus de la classe pour les mettre avec l'embryon de classe en cours de formation, et l'on continue cette opération tant que le critère choisi a priori s'améliore. Malgré des résultats intéressants cette méthode souffre d'un défaut majeur : il arrive, dans certains cas, que le niveau moyen de distance entre deux classes soit plus élevé que le niveau d'une classe qui contient leur réunion. On dit alors qu'il y a un phénomène d'inversion, et cela rend le dessin du dendrogramme impossible à réaliser sans croisement de branches.

Nous avons nous-même travaillé sur une méthode voisine (Roux, 1985, Ch. 7 ; Roux, 1991) sans parvenir à éliminer complètement ce problème des inversions. De plus, la procédure proposée, qui donne en général de bons résultats, est assez lourde en calculs et ne permet pas d'envisager des ensembles de plus de cent observations. Elle est donc inutilisable pour dépouiller de grandes enquêtes.

Il faut signaler encore la méthode de Williams et Lambert (1959) qui est bien adaptée à ce problème. En effet elle permet de traiter des tableaux de

variables qualitatives (comme par exemple les réponses à un questionnaire) même si ceux-ci ont un grand nombre d'individus. Ces auteurs font remarquer qu'une variable qualitative fournit, par elle même, une partition des individus ; les classes de cette partition sont formées des individus ayant fourni la même réponse à cette variable. On calcule l'ensemble des Chi-2 de contingence entre toutes les paires de variables, puis on retient la variable dont la somme des Chi-2 avec les autres est la plus élevée. On effectue alors la partition associée à cette variable. Puis on réitère le processus en restreignant le calcul des Chi-2 aux sous-ensembles formés par les classes que l'on veut scinder.

Le Chi-2 de contingence est élevé lorsque les deux variables sont très liées entre elles. La procédure ci-dessus vise à retenir la variable la plus liée à toutes les autres, au sein de la classe à scinder. Elle représente donc une sorte de compromis entre toutes les variables. L'inconvénient de cette méthode est son aspect "monothétique", c'est à dire que tous les individus d'une sous-classe répondent pareillement à la variable associée à la partition engendrant cette sous-classe. Dans la pratique il paraît plus raisonnable de faire des classes dans lesquelles les individus répondent "approximativement" de la même façon. L'avantage de la méthode est également son aspect "monothétique" qui permet d'associer à tout noeud de la hiérarchie obtenue le nom de la variable, ou question, "responsable" de ce noeud, facilitant ainsi l'interprétation du dendrogramme final.

5.2.2 *Classification non hiérarchique*

Pour obtenir efficacement une partition des données, la méthode la plus populaire est du type "Agrégation autour de centres variables" (Benzécri, 1973, Ch. 9) qui présente un certain nombre de variantes. Le principe de ces méthodes se résume comme suit. L'utilisateur choisit le nombre de classes à obtenir et une partition initiale. Cette partition initiale peut être tirée au hasard ou bien représenter une répartition plausible des individus en groupes obtenue empiriquement ou par d'autres méthodes. L'algorithme réitère alors deux phases successives qu'on peut appeler "Recentrage" et "Réaffectation". La phase Recentrage consiste à calculer le centre de gravité, ou point moyen, de chaque groupe, tandis que la phase Réaffectation permet de placer les individus dans le groupe dont le centre de gravité est le plus proche. Les séquences Recentrage/Réaffectation sont répétées jusqu'à obtenir la stabilité des classes, c'est à dire lorsque l'une de ces séquences ne produit plus de modification dans la composition des classes d'individus.

Dans la variante appelée K-means (Mac Queen, 1967) on effectue un recentrage dès qu'un objet change de classe. Dans la variante des "Nuées dynamiques" (Forgy, 1965 ; Diday, 1971) on effectue un recentrage global

après que tous les individus aient été réaffectés. Enfin dans la version nommée "Isodata" (Ball et Hall, 1965) un certain nombre de contraintes sont imposées pour empêcher la formation de classes d'effectifs trop faibles ou de diamètre trop grand.

On peut également envisager une variante où le point moyen, qui n'a pas toujours grand sens, est remplacé par quelques individus-types, ou "étalons" (Diday, 1971), formant en quelque sorte le noyau de la classe correspondante.

Les autres individus, ainsi que les étalons eux-mêmes, sont alors réaffectés en fonction de leurs distances moyennes aux individus-types de chaque classe.

Dans le cas où les variables sont quantitatives, et où le point moyen d'une classe est son centre de gravité, on peut montrer que la méthode des Nuées dynamiques tend à rendre maximum le moment d'inertie interclasse. Ajoutons que, dans ce cas, la méthode permet de traiter aisément de vastes tableaux de données, surtout si ceux-ci comportent peu de variables mais un grand nombre d'individus.

En effet, comme dans le cas de la construction hiérarchique du moment d'ordre deux, on peut travailler directement sur le tableau rectangulaire des données, sans avoir à gérer la matrice des distances interindividuelles. De plus la méthode est rapide puisque les seules distances à calculer sont entre les individus et un petit nombre de centres de gravité, qui peuvent également être stockés en mémoire centrale de l'ordinateur.

Ces avantages ont cependant une contre-partie défavorable : les résultats dépendent largement du choix de la partition initiale. Selon E. Diday il est cependant possible de tirer parti de cette situation.

On peut, en effet, pratiquer un certain nombre de tirages aléatoires pour créer les partitions initiales, puis réitérer la technique en partant de chacune de ces partitions. Bien entendu les classes obtenues vont être différentes d'un essai à l'autre, mais une analyse complémentaire permet de découvrir qu'un certain nombre d'individus sont toujours "sortis" dans la même classe finale, alors que d'autres ont des affectations finales qui fluctuent au cours des différents essais.

Les premiers constituent ce que l'on appelle les "Formes fortes" ; ce sont des groupes bien homogènes puisqu'ils apparaissent dans tous les cas, et leur nombre ne dépend pas beaucoup du nombre de classes choisi au départ pour la construction de la partition. Les autres individus, parfois appelés "Formes faibles", sont en général des individus isolés ou intermédiaires entre les formes fortes.

5.3 Traitement des grands tableaux d'enquêtes

Dans les paragraphes ci-dessus, qui décrivent les méthodes classiques, on a vu que certaines d'entre elles se prêtaient assez bien au dépouillement de grands tableaux d'enquêtes. Ce sont la Construction du moment d'ordre deux, pour les méthodes hiérarchiques et les techniques de Réaffectation/Recentrage pour les méthodes de partitionnement. Ces deux techniques ont pour point commun de travailler sur les centres de gravité des groupes en cours de formation, et de viser l'optimisation du moment interclasse. Malheureusement ce type de calculs s'applique uniquement à des données quantitatives, et n'est, de ce fait, que peu utilisable pour les enquêtes.

Dans le présent paragraphe on décrit d'abord une stratégie simple pour adapter les méthodes ci-dessus aux variables qualitatives. Puis on montre comment obtenir une partition à partir d'un arbre hiérarchique. Enfin on expose des méthodes spécialement élaborées pour le traitement de grands tableaux, dans lesquelles on accepte des résultats plus grossiers, pourvu qu'on puisse efficacement traiter plusieurs milliers d'individus.

Disons tout de suite qu'en pareil cas la recherche d'une hiérarchie complète sur l'ensemble des individus n'a pas grand intérêt. D'une part l'interprétation d'une telle hiérarchie, s'étendant, peut-être, sur plusieurs mètres de papier, est malaisée, d'autre part l'objectif poursuivi est plus souvent d'obtenir une partition. En effet il sera en général plus facile de raisonner sur un petit nombre de sous-échantillons, quitte à traiter chacun d'eux par des méthodes plus fines pour approfondir les résultats.

5.3.1 La stratégie "analyse factorielle + classification"

Les données d'enquêtes se présentent le plus souvent sous forme de questions ayant un nombre limité de modalités de réponse. On sait que de telles données se prêtent bien à l'analyse factorielle des correspondances multiples, qui apporte par elle-même de précieux renseignements sur la structure des données et leur interprétation. Mais on peut aussi utiliser l'analyse factorielle comme une étape préalable, ou prétraitement, avant une classification. En effet, on peut prendre les coordonnées des individus sur les premiers facteurs obtenus comme de nouvelles variables formant, en quelque sorte, un résumé du tableau initial. Ces nouvelles variables peuvent être considérées comme quantitatives et peuvent donc être introduites comme données pour les calculs de classification portant sur les centres de gravité.

La difficulté d'une telle stratégie réside dans le choix du nombre d'axes factoriels à retenir. On se base, pour cela, sur l'examen de la courbe de

décroissance des pourcentages d'inertie expliquée. Il faut éviter de conserver un axe et de rejeter l'axe suivant si leurs inerties expliquées sont voisines. En revanche une forte décroissance des inerties expliquées suivie d'une courbe en pente faible sera un bon indicateur du nombre d'axes à retenir. Rappelons que la valeur absolue de ce pourcentage d'inertie n'a pas d'intérêt par elle-même car elle dépend fortement de la dimension du tableau des données. On consultera avec profit les indications données par Lebart et al. (1977) sur cette question.

En revanche cette stratégie présente de nombreux avantages. Tout d'abord les axes de l'analyse factorielle sont très stables relativement à l'échantillonnage, ce qui n'est pas le cas des classifications (la suppression ou l'adjonction d'individus peuvent changer notoirement l'aspect d'une hiérarchie ou d'une partition). Le prétraitement par l'analyse factorielle permet de remédier partiellement à ce problème.

Le fait d'abandonner une partie de l'information initiale en ne conservant qu'un nombre restreint d'axes factoriels, loin d'être un inconvénient, peut se révéler avantageux en éliminant des fluctuations aléatoires pouvant masquer les phénomènes importants. L'analyse factorielle agit alors comme un filtre préservant l'information utile.

Enfin d'un point de vue pratique, même s'il occasionne un surcroît de calculs, le prétraitement par analyse factorielle, permet ensuite une accélération des calculs de classification, ceux-ci portant sur un tableau plus petit. Il permet d'utiliser les algorithmes basés sur l'inertie intra/inter-classe, et il fournit un point de vue original sur les données. Ce point de vue peut, et doit, être confronté avec les résultats de classification pour une meilleure interprétation globale.

5.3.2 Obtention d'une partition à partir d'une hiérarchie.

Une autre stratégie intéressante consiste à couper les branches d'un arbre hiérarchique à un certain niveau, puis de ne conserver de cet arbre que les groupes déterminés par la partie de l'arbre inférieure à ce niveau (cf. Figure 3).

Il peut paraître compliqué de passer par une construction hiérarchique pour n'en conserver qu'une partition, mais, comme ces méthodes sont rapides, cela n'alourdit pas excessivement les calculs.

De plus on peut utiliser la partition ainsi obtenue comme partition initiale dans un programme d'agrégation autour de centres variables. Comme celui-ci ne peut qu'améliorer la partition qu'on lui fournit, le résultat est

généralement de très bonne qualité. Pour des données issues d'un questionnaire nous recommandons alors la stratégie suivante, un peu compliquée mais donnant d'excellents résultats :

Analyse factorielle des correspondances multiples
+ Construction hiérarchique du moment d'ordre deux
+ Troncature de la hiérarchie
+ Agrégation autour de centres variables

Elle nécessite deux interventions manuelles ; l'une pour le choix du nombre d'axes factoriels à retenir après l'analyse factorielle, l'autre pour déterminer le niveau de troncature de la hiérarchie.

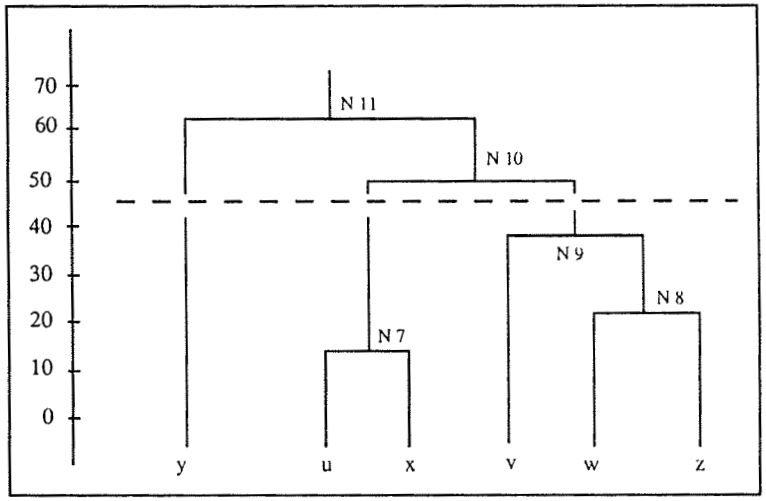

Figure 3

La troncature de l'arbre hiérarchique de la Fig. 1 au niveau 42 fait apparaître 3 classes : {y}, {u, x}, {v, w, z}

5.3.3 Obtention directe d'une partition

Trois familles de méthodes se proposent d'obtenir directement une partition. La première relève de ce qu'on appelle la recherche des zones de densité élevée, ou recherche des modes. Elle considère les individus comme représentés par des points d'un espace vectoriel ; nous avons vu que, grâce à l'analyse factorielle, ceci n'est pas un obstacle.

On attribue à chaque point un coefficient de densité, qui est fonction du nombre et de la distance de ses voisins, et les points les plus denses seront les embryons des groupes cherchés.

Dans la deuxième famille de méthodes on divise l'espace en petits compartiments, aux cloisons parallèles aux axes de référence. Puis ce compartimentage est utilisé pour réduire considérablement le nombre de distances interindividuelles à calculer dans la recherche de classes homogènes.

La troisième famille de méthodes consiste à réaliser une partition sur un sous-échantillon, puis à répartir le reste des individus entre les classes formées, en fonction des distances moyennes à ces classes.

5.3.3.1 Coefficients de densité.

Si l'on se fixe a priori un rayon R, on peut dire que la densité en x est le nombre de points inclus dans la boule B(x, R), de centre x et de rayon R. Ou bien, si l'on se donne un effectif K, on peut définir la densité comme l'inverse du rayon R', de la boule B(x, R') qui contient K observations. Une fois choisie la mesure de la densité, il faut encore choisir un autre paramètre D, qui indique la distance minimum entre un groupe et un individu pour que ce dernier puisse constituer l'embryon d'un nouveau groupe. On range ensuite les individus par ordre de densités décroissantes, puis on les examine successivement dans cet ordre. Le premier individu de la liste forme le début d'un premier groupe, l'examen des individus suivants peut donner lieu à trois décisions différentes :

1) L'individu est à une distance inférieure à D de l'un des groupes en cours de formation. On l'agrège alors à ce groupe.

2) L'individu est à une distance supérieure à D de tous les groupes en cours de formation. Dans ce cas il initialise un nouveau groupe.

3) L'individu est à une distance inférieure à D de plusieurs groupes distincts. Les deux groupes sont alors fusionnés et l'individu est agrégé au nouveau groupe.

Ce schéma général admet de nombreuses variantes qui visent à limiter le nombre de groupes ou bien à imposer un effectif minimum aux nouveaux groupes (Wishart, 1969 ; Carmichael et Sneath, 1969 ; Fortin 1975).

Remarquons que l'on peut aisément s'affranchir du choix arbitraire du paramètre D, ou du paramètre R.

En effet on peut choisir une fonction f décroissante, et décider de mesurer la densité au point i, par la formule :

$$\text{Dens (i)} = \sum_{j \in J} f(d_{ij})$$

Ainsi, si on prend pour fonction f une exponentielle décroissante, les points très éloignés de i n'auront qu'une contribution minime à la densité :

$$\text{Dens (i)} = \sum_{j \in J} \text{Exp}[-k\ d_{ij}]$$

où k serait un paramètre positif réglant la rapidité de la décroissance de la fonction.

5.3.3.2 Découpage de l'espace en compartiments, ou pavés

L'algorithme C.C.C (Clustering by Connected Components) de Hansen et Lehert (1980) est très rapide et permet donc le traitement de très grands jeux de données. Chaque dimension de l'espace vectoriel est découpée en classes d'égale amplitude A. De sorte que la partie de l'espace vectoriel qui contient les points à classer se trouve compartimentée en un certain nombre d'hypercubes de côté A. On dit aussi que l'on a réalisé un "pavage" de l'espace ; chaque hypercube étant encore appelé un "pavé". Le comptage du nombre de points inclus dans chaque pavé donne une autre façon d'apprécier la densité, et l'on pourrait appliquer l'un des algorithmes du paragraphe précédent pour agréger les pavés entre eux ou initialiser des groupes de pavés.

Mais ce n'est pas ce que font Hansen et Lehert. Ils considèrent les individus comme les sommets d'un graphe, sommets qui sont tous reliés deux à deux (Graphe complet) par des arêtes dont la longueur est égale à la distance entre les individus correspondants. Ce graphe en lui-même est sans intérêt, mais on peut en chercher un sous-graphe plus simple, et donc mieux représentable, en éliminant les arêtes de longueur supérieure à A.

Hansen et Lehert ont pour objectif de déterminer ce que l'on appelle les composantes connexes de ce sous-graphe, c'est à dire les sous-ensembles d'individus reliés entre eux par des chaînes dont chaque arête est de longueur inférieure à A. Ils utilisent le découpage de l'espace en hypercubes pour éviter d'avoir à calculer de très nombreuses distances interindividuelles.

On appelle H(i) le pavé contenant un individu i et V(i) son voisinage. Ce voisinage est constitué de H(i) et de l'ensemble des hypercubes, ou pavés, adjacents de H(i), c'est-à-dire ceux dont les arêtes sont adjacentes, ou confondues, avec les arêtes de H(i). Si on examine deux individus i et i' deux cas peuvent se présenter. Ou bien i' appartient au voisinage V(i) ou bien il lui est extérieur ; dans ce deuxième cas on peut être certain que sa distance à i est supérieure à A, dans le premier cas, au contraire, il faut effectivement calculer cette distance pour s'en assurer.

Pour décrire le déroulement de l'algorithme CCC on appelle "individu libre" un individu qui n'est pas encore relié à un autre (dans le sous-graphe) ; on appelle "individu fixé" un individu qui est connecté à un autre, enfin on dit qu'un individu est "validé" si tous ses voisins distants d'au plus A ont bien été fixés. Une fois réalisé le pavage de l'espace et la détermination des individus appartenant à chaque pavé, l'algorithme se compose de deux boucles imbriquées :

- *Boucle 1 :*
  Tant qu'il reste des individus libres, choisir un individu libre k et dresser la liste de tous les individus inclus dans la boule B(k, A), qui sont déclarés fixés.
  - *Boucle 2 :*
    Tant qu'il reste des individus fixés, choisir un individu fixé l, et dresser la liste des individus libres inclus dans la boule B(l, A) ; ces individus deviennent fixés à leur tour.
  - *Fin-boucle 2 :*
    L'ensemble des individus fixés, y compris k, deviennent validés et forment une composante connexe, c'est-à-dire une classe de la partition cherchée.
- *Fin-boucle 1.*

La recherche des individus des boules B(k, A) et B(l, A) est très rapide car seuls les points appartenants aux voisinages de k et de l, respectivement, sont examinés. L'algorithme est encore accéléré par un rangement préalable des individus harmonisé avec l'ordre des pavés, de sorte que la recherche des individus d'un voisinage ne nécessite pas le balayage de l'ensemble de tous les individus.

Cet algorithme présente deux inconvénients.
Le premier est qu'il faut choisir la dimension A des pavés, qui sert également de seuil aux distances à conserver. Mais la rapidité de l'algorithme fait que l'on peut le répéter plusieurs fois avec des valeurs de A différentes.

Le second inconvénient est que les classes obtenues coïncident avec celles que l'on obtiendrait en tronquant la hiérarchie obtenue par l'agrégation selon le saut minimum. Or on sait que le défaut de cette méthode est l'effet de chaîne (§ 5.2.1.1) qui fournit des classes dont les individus extrêmes peuvent être très différents, pourvu qu'il existe une série d'intermédiaires à faible distance les uns des autres. On peut montrer, cependant, que la partition obtenue a la propriété d'être de "séparation" maximum, la séparation d'une partition étant la distance minimum qui sépare deux classes quelconques.

5.3.3.3 Utilisation d'un sous échantillon

La méthode "CLARA" de Kaufmann et Rousseuw (1986) commence par tirer au hasard, un sous-échantillon de l'ensemble des données. On réalise ensuite une partition de ce sous-échantillon par une méthode apparentée à une recherche de zones denses (§ 5.3.3.1), mais dans laquelle il n'est pas besoin de choisir un paramètre arbitraire. Lorsque ceci est terminé on affecte les individus mis à l'écart dans le tirage au hasard à l'un ou l'autre des groupes obtenus en fonction des distances moyennes entre ces individus et ces groupes. Une évaluation de la partition globale obtenue est calculée selon le critère de la distance moyenne entre les individus et le représentant de la classe à laquelle ils appartiennent. L'algorithme CLARA se poursuit en répétant plusieurs fois les opérations ci-dessus et en ne conservant, comme résultat définitif, que la meilleure des partitions globales. Un certain nombre de variantes de cette méthode existent comme celle de Steinhausen et Langer (1977).

5.4 Aides pour l'interprétation d'une partition

Une fois obtenue une partition des individus en un certain nombre de classes, l'utilisateur désire connaître ce qui a motivé le rapprochement des individus appartenant à une même classe, et par voie de conséquence, ce qui les différencie des individus des autres classes. Autrement dit on désire savoir quelles sont les variables, ou questions, "responsables" des classes obtenues. Un certain nombre de calculs complémentaires sont donc les bienvenus pour permettre une meilleure compréhension des résultats.

5.4.1 Informations complémentaires sur les individus

Tout bon programme de calcul de partition devrait fournir les renseignements suivants, faciles à calculer. Pour chaque classe :

- l'effectif de la classe,
- son diamètre (distance entre les 2 points les plus éloignés),
- la séparation (distance minimum entre la classe considérée et la classe qui en est la plus proche) et le numéro de la classe la plus proche,
- le numéro de l'individu le plus central de la classe (celui dont la distance moyenne aux autres est la plus petite) et la liste de ses réponses aux différentes questions (ou variables),
- les numéros des individus périphériques, avec le numéro de la classe la plus proche de ces individus.

5.4.2 *Rôle des variables quantitatives*

L'idée de base pour déterminer le rôle des variables quantitatives est d'utiliser l'équation (4), ou son écriture détaillée l'équation (6), du paragraphe 5.2.1.2, qui dit que la somme des carrés des écarts totaux se décompose en la somme des carrés des écarts intra-classes, ou résiduels, et la somme des carrés des écarts interclasses, ou factoriels.

Dans l'équation (6) on fait intervenir des carrés de distances entre individus et centres de gravité (désignés ci-après par la lettre z). Nous supposerons dans la suite que les distances sont calculées selon la formule euclidienne usuelle décrite par la formule (9), que nous rappelons ici :

$$d^2(x,y) = \sum_{j \in J} (x_j - y_j)^2$$

En utilisant cette expression dans la formule (6) on obtient :

$$\sum_{i \in I} \sum_{j \in J} (x_{ij} - z_j)^2 = \sum_{p \in P} \sum_{i \in p} \sum_{j \in J} (x_{ij} - z_{pj})^2 + \sum_{p \in P} \sum_{j \in J} n_p (z_{pj} - z_j)^2 \qquad (12)$$

expression dans laquelle on peut mettre en avant les sommations portant sur l'indice j. Cette opération permet de mettre en relief le rôle des variables qui peut se quantifier dans le rapport suivant, que nous appelons "contribution de la variable j au groupe p" :

$$CV(j,p) = (z_{pj} - z_j)^2 \; / \; \sum_{j \in J} (z_{pj} - z_j)^2 \qquad (13)$$

Ce rapport donne le rôle de la variable j dans l'éloignement du groupe p du centre de gravité général. C'est donc bien la part de j dans ce qui fait l'originalité du groupe p. La somme de ces contributions sur l'ensemble des variables est égale à 1.

Un autre ensemble valeurs numériques utiles est donné par la formule suivante :

$$CG(p,j) = n_p (z_{pj} - z_j)^2 \; / \; \sum_{p \in P} n_p (z_{pj} - z_j)^2 \qquad (14)$$

Nous les appelons "contributions des groupes aux variables". Elles indiquent comment l'inertie de la variable se décompose suivant les différents groupes. Les valeurs fortes marquent les groupes les mieux caractérisés par cette variable. La somme de ces valeurs sur l'ensemble des groupes est égale à 1.

5.4.3 *Rôle des variables qualitatives*

Quand les données sont qualitatives les notions ci-dessus ne s'appliquent plus, mais on peut utiliser la mesure usuelle de coefficient d'association représentée par le Chi-2 de contingence. En effet le résultat de la classification est en général une partition en K classes, qui peut être considérée comme une nouvelle variable à K modalités. Les variables les plus liées à la partition seront celles pour lesquelles le Chi-2 de contingence avec la variable "Partition" est le plus fort.

Supposons que l'on ait une question dont l'ensemble des modalités de réponse est désigné par Q, et que l'on appelle P l'ensemble des classes de la partition étudiée. La formule du Chi-2 s'écrit :

$$\chi^2(P,Q) = \sum_{p \in P} \sum_{q \in Q} (n_{pq} - e_{pq})^2 / e_{pq} \qquad (15)$$

où n_{pq} désigne le nombre d'individus de la classe p de la partition ayant répondu q à la question considérée ; e_{pq} désigne l'effectif théorique de la case correspondante, qui se calcule comme suit :

$$e_{pq} = (n_{p.}\ n_{.q}) \; / \; n$$

Dans cette formule n désigne l'effectif total, tandis que $n_{p.}$ représente l'effectif de la classe p et $n_{.q}$ le nombre d'individus ayant donné la réponse q à la question considérée.

Dans l'expression (15) ci-dessus on peut étudier les valeurs des termes qui entrent dans la fabrication du Chi-2. Elles décrivent le rôle des modalités de réponses dans l'élaboration de la partition finale. La quantité :

$$CV(p,q) = (n_{pq} - e_{pq})^2 \; / \; e_{pq} \qquad (16)$$

peut être considérée comme la contribution de la réponse q à la classe p. Pour une classe p fixée, les valeurs les plus élevées sont associées aux réponses caractéristiques de cette classe.

Si toutes les questions ont le même nombre de réponses possibles, on pourra comparer les valeurs $\chi^2(P,Q)$ entre elles, les valeurs les plus élevées correspondent alors aux variables les plus caractéristiques de la partition.

Dans le cas où les questions ont des nombres inégaux de réponses, il faut utiliser ces quantités avec circonspection car la valeur du Chi-2 est très sensible au nombre de classes des variables prises en compte. Dans ce cas on remplacera avantageusement le Chi-2 par le coefficient de Cramer (1946) qui n'est autre que le Chi-2 divisé par le produit nv où v est le plus

petit des nombres de catégories dans les deux variables étudiées. On peut montrer que ce coefficient a l'avantage d'être compris entre 0 et 1.

5.5 Conclusion

Nous avons vu que les méthodes traditionnelles de construction hiérarchique étaient d'un intérêt limité pour le traitement des enquêtes. Cependant la construction ascendante du moment d'ordre deux, ou méthode de Ward, s'allie efficacement avec les constructions de partition du type "Réaffectation/Recentrage" en fournissant une partition initiale de bonne qualité. L'exigence de variables quantitatives pour ces méthodes peut être satisfaite grâce à un traitement préalable par l'analyse factorielle. Par ailleurs un certain nombre de techniques ont été mises au point pour traiter directement de vastes ensembles d'individus.

Dans tous les cas le résultat final se présentera sous la forme d'une partition, car une hiérarchie sur des centaines, voire des milliers d'individus, serait inexploitable. Même si c'est bien le but recherché, une partition est un résultat assez pauvre, mais il peut être considérablement enrichi par des calculs complémentaires, notamment ceux qui décrivent les contributions des variables aux classes de la partition obtenue.

6

PREPARATION DES TABLEAUX POUR L'ANALYSE DES DONNEES : LE CODAGE DES VARIABLES

Yvette Grelet

Laboratoire d'Economie Sociale
Centre d'Etudes et de Recherches sur les Qualifications
Paris

L'analyse d'un ensemble de données comprend trois grandes étapes : la préparation du tableau soumis à l'analyse, la chaîne des traitements elle-même, et la phase d'interprétation ; ces trois étapes n'étant dissociables que par l'ordre de leur enchaînement, mais non dans leur conception. Nous nous intéresserons ici à la première de ces étapes.

Construire un tableau c'est faire le choix de l'ensemble des lignes, de l'ensemble des colonnes et des nombres qui, à la croisée des lignes et des colonnes, auront en charge de traduire au mieux la réalité qu'on s'est fixée comme objet d'étude. C'est en quelque sorte choisir un angle de prise de vue, qui oriente le regard vers la chose à voir. Comme la première prise est rarement la bonne, *on procédera par itérations*, les résultats de l'analyse conduisant à retourner aux données pour supprimer une ligne, regrouper deux colonnes, tenter un nouveau codage. Ces essais successifs qui peuvent paraître laborieux à l'utilisateur pressé, et relever du tâtonnement, sont d'une grande utilité : ils permettent d'affiner l'analyse et d'en valider les résultats.

6.1 Du questionnaire au tableau de données

Prenons pour fixer les idées l'une des enquêtes du CEREQ sur l'insertion des jeunes à la sortie du système éducatif. Le champ éducatif est couvert, avec une périodicité de quatre ans, par vagues d'enquête découpées selon le niveau de formation. Il s'agira ici de la cohorte des garçons et filles ayant quitté l'école en 1986 au niveau de l'enseignement secondaire général ou technique dont un échantillon de 10 500 jeunes a été interrogé en décembre 1989.