Chapitre 4

La classification automatique des formes et des textes

Les méthodes de classification automatique, constituent à côté des méthodes factorielles, la seconde grande famille de techniques d'analyse des données. Ces méthodes permettent de représenter les proximités entre les éléments d'un tableau lexical (lignes ou colonnes) par des regroupements ou classes.

Cet ensemble se divise lui-même en deux familles principales :

a) Les méthodes de *classification hiérarchique*, qui permettent d'obtenir à partir d'un ensemble d'éléments décrits par des variables (ou dont on connaît les distances deux à deux), une hiérarchie de classes partiellement emboîtées les unes dans les autres.

b) Les méthodes de *partitionnement*, ou de classification directe, qui produisent de simples découpages ou partitions de la population étudiée, sans passer par l'intermédiaire d'une hiérarchie. Ces dernières sont mieux adaptées aux très grands ensembles de données (plusieurs milliers d'individus à classer).

Les méthodes correspondant aux deux familles peuvent être combinées en une approche mixte qui sera évoquée au paragraphe 4.3.

Les résultats fournis par les méthodes de classification se révèlent, dans la pratique, des compléments indispensables aux résultats fournis par l'analyse des correspondances qui ont été décrits au chapitre précédent.

En effet, dans l'étude des tableaux lexicaux, l'analyse des correspondances permet avant tout de dégager de grands traits structuraux qui portent à la fois sur les deux ensembles mis en correspondance. Cependant, lorsque le nombre des éléments représentés est important, il devient délicat, dans la pratique, d'apprécier leurs positions réciproques au seul vu des résultats graphiques. Dans ce cas en effet, les plans factoriels deviennent peu lisibles en raison du grand nombre de ces éléments. Il en va de même, dès que le texte du corpus analysé est un peu long même si le nombre des parties est relativement

restreint. En effet, le nombre des formes croît alors inévitablement de manière importante, même si on restreint l'analyse aux formes les plus fréquentes.

Une seconde raison qui motive l'emploi conjoint des méthodes factorielles et de la classification tient à l'enrichissement dimensionnel des représentations obtenues. Les regroupements obtenus se font en effet à partir de distances calculées *dans tout l'espace*, et non seulement dans le ou les premiers plans factoriels. Ils seront donc de nature à corriger certaines des déformations inhérentes à la projection sur un espace de faible dimension.

La dernière raison de cette complémentarité est pratique : il est plus facile de faire décrire automatiquement des classes par l'ordinateur, que de faire décrire un espace continu, même si la dimension de celui-ci a été fortement réduite. Un exemple tentera de montrer l'intérêt de cette "dissection" de l'espace.

Dans ce chapitre on se limitera à la description et à l'utilisation de méthodes de classification automatique qui utilisent la même métrique que l'analyse des correspondances, la métrique du chi-2 définie au chapitre précédent, de façon à assurer une bonne compatibilité des résultats.

Après un bref rappel sur les méthodes elles-mêmes (4.1), la classification sera utilisée pour décrire tour à tour les proximités sur les formes et sur les parties d'un corpus (4.2). On montrera ensuite quel rôle elle peut jouer dans le cas de fichiers de type "enquête" (4.3).

4.1 Rappel sur la classification hiérarchique

Comme l'analyse des correspondances, la classification hiérarchique s'applique aux tableaux à double entrée tels les tableaux lexicaux décrits dans les chapitres précédents. On peut soumettre à la classification soit l'ensemble des colonnes du tableau (qui correspondent la plupart du temps aux différentes parties d'un corpus) soit celui des lignes de ce même tableau (qui correspondent en général aux formes et parfois aux segments du même corpus).

Le principe de l'agrégation hiérarchisée est très simple dans son fondement. On part d'un ensemble de n éléments (que l'on appellera ici des éléments de base ou encore *éléments terminaux*) dont chacun possède un poids, et entre lesquels on a calculé des distances (il y a alors $n(n-1)/2$ distances entre les différents couples possibles). On commence par agréger les deux éléments les plus proches. Le couple ainsi agrégé constitue alors un *nouvel élément* dont on peut recalculer à la fois le poids et les distances à chacun des éléments qu'il

reste à classer.[1] A l'issue de cette étape, le problème se trouve ramené à celui de la classification de *n-1* éléments. On agrège à nouveau les deux éléments les plus proches, et l'on réitère ce processus (*n-1* fois au total) jusqu'à épuisement de l'ensemble des éléments. L'ultime et *(n-1)éme* opération regroupe l'ensemble des éléments au sein d'une classe unique.

Chacun des regroupements effectués en suivant cette méthode s'appelle un *noeud*. L'ensemble des éléments terminaux rassemblés dans un noeud s'appelle une *classe***.** Les deux éléments (ou groupes d'éléments) agrégés, l'*aîné* et le *benjamin* de ce noeud.

Ce principe est celui de la classification *ascendante* hiérarchique, famille de méthodes la plus répandue. Il faut noter qu'il existe aussi de méthodes *descendantes*, qui opèrent par dichotomies successives de toute la population.[2]

4.1.1 Le dendrogramme

La classification ainsi obtenue peut être représentée de plusieurs manières différentes. La représentation sous forme d'arbre hiérarchique ou *dendrogramme* constitue sans doute la représentation la plus parlante.

Les regroupements effectués à chaque pas de l'algorithme de classification hiérarchique rassemblent des éléments qui sont plus ou moins proches entre eux. Plus on avance dans le regroupement (plus on se rapproche du sommet de l'arbre), plus le nombre de points déjà agrégés est important et plus la distance minimale entre les classes qu'il reste à agréger est importante. On peut associer à chacun des noeuds de l'arbre cette "plus petite distance".

Cette manière de procéder permet de mettre en évidence, comme sur la figure 4.1 ci-dessous, les noeuds qui correspondent à des augmentations importantes de cette distance minimale et qui rassemblent donc des composants nettement moins homogènes que leur réunion.
On lit sur la figure 4.1 que les éléments C et D, très proches, se sont agrégés dès la première itération, et que les deux blocs [A, B, C, D] d'une part, et [E, F, G, H] d'autre part, qui s'agrègent en dernier, constituent probablement les deux classes d'une bonne partition de la population (i.e. des 8 éléments terminaux) vis-à-vis de la distance utilisée.

---

[1] Dans la pratique il existe un grand nombre de façons de procéder qui correspondent à cette définition, ce qui explique la grande variété des méthodes de classification automatique. Comme nous l'avons signalé plus haut, nous n'utiliserons ici qu'une seule de ces variantes, basée sur l'utilisation de la distance du chi-2 et la maximisation du moment d'ordre 2 d'une partition (Benzécri, 1973), appelée parfois critère de Ward généralisé.

[2] Une méthode de ce type a été préconisée par Reinert (1983) dans le domaine textuel pour classer des *unités de contexte*

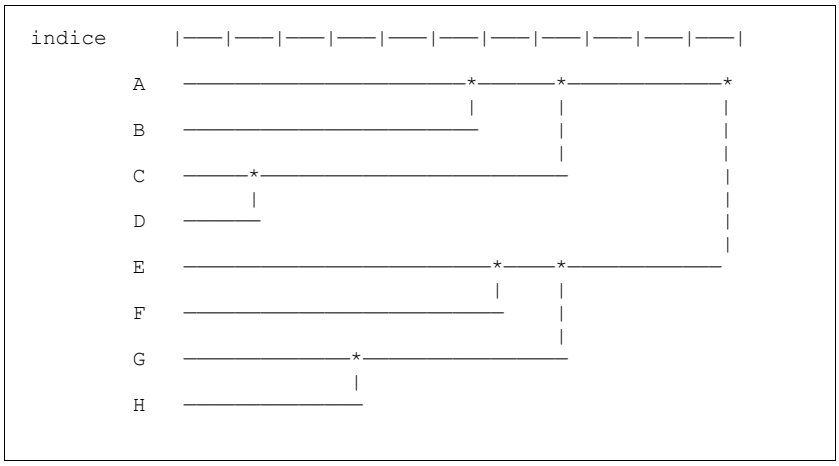

Figure 4.1

Dendrogramme représentant une classification sur un ensemble de 8 éléments

La représentation sous forme de dendrogramme d'une classification hiérarchique matérialise bien le fait que les classes formées au cours du processus de classification constituent une *hiérarchie indicée* de classes partiellement emboîtées les unes dans les autres.

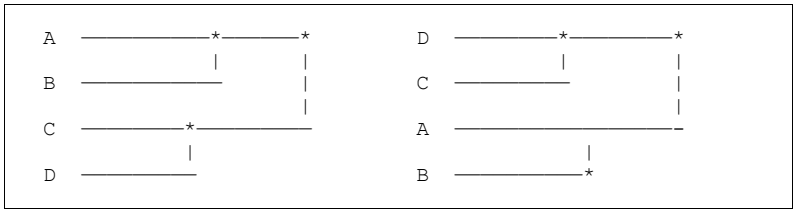

Figure 4.2

Dendrogrammes équivalents

L'interprétation de cette hiérarchie s'appuie sur l'analyse des seules distances entre éléments ou classes faisant l'objet d'un même noeud. En effet, pour une classification construite sur un ensemble d'éléments [A, B, C, D], la représentation sous forme de dendrogramme n'est pas unique
puisque chaque permutation de l'aîné et du benjamin d'un même noeud amène une représentation équivalente de la hiérarchie des classes.
Ainsi, les deux représentations de la figure 4.2 correspondent-elle à une même classification des quatre éléments en une hiérarchie comportant, en dehors des éléments terminaux, les trois classes [C,D], [A,B], [A,B,C,D].

4.1.2 Coupures du dendrogramme

Si le nombre des éléments à classer est important, la représentation complète de l'arbre de classification devient difficile à étudier. Une solution pratique consiste à définir un *niveau de coupure* de l'arbre (qui correspond précisément à un intervalle entre deux des itérations du processus de classification ascendante). Une telle coupure permet de considérer une classification résumée aux seules classes supérieures de la hiérarchie.

On peut définir la coupure du dendrogramme en déterminant à l'avance le nombre des classes dans lesquelles on désire répartir l'ensemble des éléments à classer, ou encore, ce qui revient au même, le nombre de noeuds supérieurs que l'on retiendra pour la classification (le nombre des classes retenues est alors égal au nombre des noeuds augmenté d'une unité).

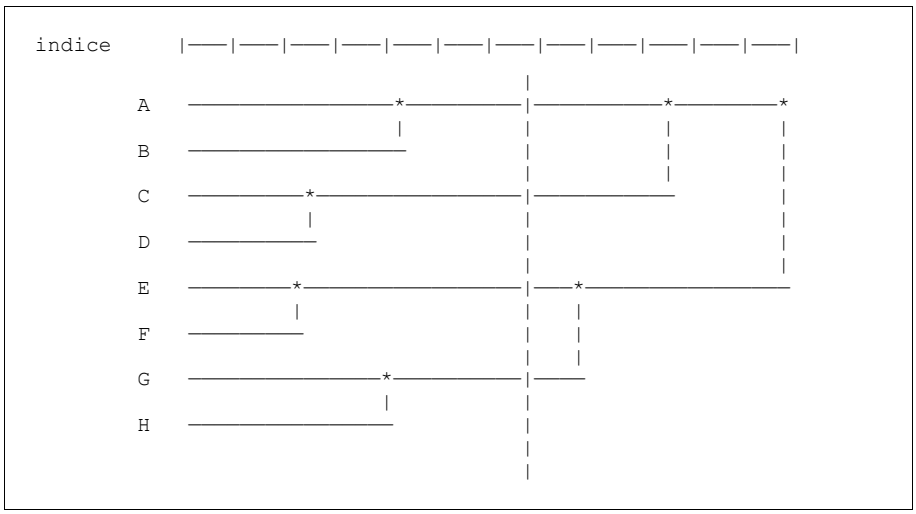

Figure 4.3

Coupure d'un dendrogramme réalisant 4 classes

La coupure pratiquée sur le dendrogramme représenté sur la figure 4.3 opère sur l'ensemble des éléments de départ une partition[1] en quatre classes [A, B], [C, D], [E, F], [G, H], dont la réunion correspond à l'ensemble soumis à la classification.

Chacune des classes est décrite exhaustivement par la liste des éléments qu'elle contient. Le choix d'une coupure dans l'arbre de classification est une opération qui doit en premier lieu s'appuyer sur les valeurs de l'indice : il faut dans la mesure du possible que cette coupure corresponde à un saut de l'indice. Avec les conventions des figures 4.1 à 4.3, les valeurs de l'indice à gauche de la coupure doivent être faibles, et celles à droite de la coupure forte : de cette

[1] Il s'agit ici de la partition de l'ensemble des éléments à classer. Cette notion, courante dans le domaine de la classification automatique doit être distinguée de celle de partition d'un corpus de textes (i.e. la division d'un corpus de textes en parties).

façon, les éléments sont proches à l'intérieur des classes définies par la coupure, et éloignés lorsqu'ils appartiennent à des classes différentes. On reviendra sur ce point au paragraphe 4.3.1 dévolu aux méthodes de classification mixtes.

4.1.3 Adjonction d'éléments supplémentaires

A partir d'une classification hiérarchique réalisée sur un ensemble d'éléments, il est possible de répartir à l'intérieur des classes de la hiérarchie ainsi créée un ensemble d'éléments supplémentaires ou illustratifs.[1]

Pour chacun des éléments supplémentaires, l'algorithme d'adjonction procède de manière particulièrement simple. On commence par rechercher parmi les éléments de l'ensemble de base celui dont l'élément supplémentaire à classer est le plus proche. On affecte ensuite cet élément supplémentaire à toutes les classes de la hiérarchie qui contiennent l'élément de base.

Cette méthode peut être utilisée, comme dans le cas de l'analyse des correspondances, pour illustrer par des données segmentales la classification obtenue à partir des formes. Ici encore, la méthode utilisée permet de classer un nombre important d'éléments supplémentaires sans perturber la classification obtenue sur l'ensemble de base.

4.1.4 Filtrage sur les premiers facteurs

Comme on l'a signalé plus haut, la classification hiérarchique se construit à partir d'une distance calculée entre les couples d'éléments de l'ensemble de base pris deux à deux. Dans le cas général, on mesure les distances entre les éléments soumis à la classification à l'aide de la distance du chi-2 entre les colonnes du tableau, distance qui est également à la base de l'analyse des correspondances.
Rappelons cependant que le but de l'analyse des correspondances est précisément l'extraction, lorsqu'elle est possible, de sous-espaces résumant au mieux l'information contenue dans le tableau de départ. L'interprétation s'appuie en général sur l'hypothèse que les facteurs qui correspondent aux valeurs propres les plus faibles ne constituent qu'un "bruit" non susceptible d'interprétation.

Si l'on suit cette logique, on peut opérer, à partir de l'ensemble des éléments de base, des classifications qui font intervenir les distances mesurées dans le seul

[1] On utilise ici cette notion dans le sens précis de la définition des éléments supplémentaires introduite au chapitre précédent.

espace des premiers facteurs, réputés les plus significatifs au sens statistique du terme.

L'opération qui transforme les distances mesurées entre les éléments de base en les réduisant à leur projection dans l'espace des premiers facteurs constitue un *filtrage* des distances sur les premiers facteurs. Elle permettra, par une réduction sélective de l'information, de classer des milliers de formes ou d'individus.[1]

Cette propriété de filtrage commune à la plupart des méthodes fondées sur la recherche d'axes principaux (décomposition aux valeurs singulières, analyse en composantes principales, analyse des correspondances simples et multiples) sera à nouveau évoquée au paragraphe 7.4 (chapitre 7). Elle est aussi utilisée en *discrimination textuelle* (chapitre 8) et en recherche documentaire dans l'approche désignée par l'expression *latent semantic analysis* (Deerwester et al., 1990 ; Bartell et al., 1992).

4.2 Classification des lignes et des colonnes d'un tableau lexical

Comme au chapitre précédent, on présentera le fonctionnement de la méthode sur un exemple de dimensions réduites, en rappelant que cet exercice pédagogique qui doit mettre en évidence le mécanisme des opérations ne parviendra pas à donner une idée exacte de la valeur heuristique de l'outil. Seront successivement examinées la classification des lignes (formes) et la classification des colonnes (parties).

4.2.1 Classification des formes

On reprendra l'exemple du chapitre 3 croisant 14 formes graphiques et 5 catégories de répondants à une enquête. La classification s'opérera sur les lignes du tableau 4.1 (lequel reproduit le tableau 3.2 du chapitre précédent). Les lignes de ce tableau sont les profils-lignes entre lesquels sont calculés les distances qui nous intéressent.
La première colonne de ce tableau contient un numéro d'ordre qui servira par la suite à décrire les regroupements. Les effectifs originaux de chaque ligne figurent dans la dernière colonne. Le tableau 4.2 décrit les étapes et les paramètres du fonctionnement de l'algorithme.

[1] Une analyse des correspondances préalable procure deux autres avantages : elle fournit une description des données complémentaire, parce que reposant sur des principes différents ; elle peut permettre des économies considérables lors des calculs de distances dans le cas de grands tableaux (en recherche documentaire par exemple).

Tableau 4.1

Rappel des profils-lignes du tableau 3.2 (chapitre 3)

		SDipl (1)	CEP (2)	BEPC (3)	Bacc (4)	Univ (5)	Total	(Eff)
1	*Argent*	26.4	33.2	16.6	15.0	8.8	100.	(193)
2	*Avenir*	16.7	28.3	24.5	23.6	6.9	100.	(318)
3	*Chômage*	25.1	39.2	17.7	14.1	3.9	100.	(283)
4	*Conjoncture*	4.5	31.8	22.7	22.7	18.2	100.	(22)
5	*Difficile*	25.9	40.7	14.8	11.1	7.4	100.	(27)
6	*Économique*	13.0	24.1	22.2	20.4	20.4	100.	(54)
7	*Égoïsme*	19.6	34.6	13.1	24.3	8.4	100.	(107)
8	*Emploi*	15.2	44.3	24.1	7.6	8.9	100.	(79)
9	*Finances*	35.7	25.0	25.0	10.7	3.6	100.	(28)
10	*Guerre*	15.4	26.9	26.9	23.1	7.7	100.	(26)
11	*Logement*	15.4	42.3	13.5	19.2	9.6	100.	(52)
12	*Peur*	15.7	28.3	23.9	23.9	8.2	100.	(159)
13	*Santé*	19.4	29.0	21.5	20.4	9.7	100.	(93)
14	*Travail*	23.2	40.4	19.2	9.3	7.9	100.	(151)
	Total	20.3	33.7	20.2	17.9	7.9	100.	(1592)

Lecture du tableau 4.2

Première étape : Il y a 14 éléments à classer.

La première ligne du tableau 4.2 indique que le premier élément artificiel (*noeud*) est obtenu par fusion des éléments 2 et 10 (*aîné* et *benjamin*) qui sont les formes *avenir* et *guerre*. Ce nouvel élément portera le numéro 15. Il sera caractérisé par le profil moyen de ses deux constituants. La valeur de l'indice correspondante (0.00008) décrit la plus petite distance correspondante, et la masse du nouvel élément (344) est bien la somme des effectifs (344 = 318 + 26).

Seconde étape : 13 éléments à classer (seconde ligne du tableau 4.2).

Les deux éléments les plus proches, qui vont constituer l'élément 16, sont les éléments 15 (*avenir* + *guerre*) et 12 (*peur*).

Troisième étape : 12 éléments à classer (troisième ligne du tableau 4.2)

L'élément 17 est ensuite formé de l'union des éléments 14 et 5 (*travail*, *difficile*), l'élément 18 rassemble le couple (*conjoncture, économique*). Puis s'unissent (*égoïsme*, *logement*), ([*avenir* + *guerre* + *peur*], et *santé*), etc.

Le processus se termine lorsqu'il ne reste plus qu'un seul élément.

Tableau 4.2

Classification hiérarchique des formes
(description des noeuds)

NUM.	AINE	BENJ	EFF.	POIDS	INDICE	HISTOGRAMME DES INDICES DE NIVEAU
15	2	10	2	344.00	.00008	*
16	15	12	3	503.00	.00018	*
17	14	5	2	178.00	.00022	*
18	6	4	2	76.00	.00061	**
19	7	11	2	159.00	.00094	***
20	16	13	4	596.00	.00107	***
21	8	17	3	257.00	.00198	******
22	9	1	2	221.00	.00219	******
23	3	22	3	504.00	.00321	*********
24	21	23	6	761.00	.00523	**************
25	20	19	6	755.00	.00654	*****************
26	25	18	8	831.00	.00890	************************
27	26	24	14	1592.00	.03091	***************************/ /*****

SOMME DES INDICES DE NIVEAU = .06206

On note que la somme des indices (0.062) est, elle aussi, égale à la somme des valeurs propres calculée au chapitre 3. Cette quantité est, on l'a vue, proportionnelle au chi-2 (ou : χ^2) calculé sur la table de contingence.

Analyse des correspondances et classification décomposent donc de deux façons différentes la même quantité (chi-2 classique), qui mesure un écart entre la situation observée et l'hypothèse d'indépendance des lignes et des colonnes de la table.

La représentation graphique de ce dendrogramme va illustrer de façon plus suggestive ce processus d'agrégation. Elle sera confrontée à la représentation graphique obtenue précédemment par analyse des correspondances, de façon à mettre en évidence l'originalité de chacun des points de vue.

La figure 4.4 ci-après représente sous la forme d'un dendrogramme les regroupements successifs du tableau 4.2 : la longueur des branches de l'arbre est proportionnelle aux valeurs de l'indice.

Lecture de la figure 4.4

On lit sur cette figure que les formes *peur, guerre, avenir* sont agrégées très tôt, mais également que le "paquet" *(peur, guerre, avenir, santé)* ne rejoint le couple *(logement, égoïsme)* que beaucoup plus tard.

La comparaison avec la figure 3.1 du chapitre précédent est intéressante. Les deux branches principales du dendrogramme, opposant les 8 premières lignes (de *peur* à *économique*) aux six dernières (de *difficile* à *chômage*) traduisent la principale opposition visible sur le premier axe factoriel (horizontal).

Les principaux regroupements observables dans le plan factoriel de la figure 3.1, sont ceux que décrit le processus d'agrégation hiérarchique, avec cependant quelques différences qui méritent notre attention.

Les points *santé* et *égoïsme* sont proches sur la figure 3.1 (il est vrai qu'ils sont au voisinage de l'origine, et que les proximités en analyse des correspondances sont d'autant plus fiables que les points occupent des positions périphériques...).

Le dendrogramme montre au contraire que le point *santé* est plus proche de la constellation (*peur, guerre, avenir*) que du point *égoïsme*. De la même façon, ce dendrogramme nous montre que le point *égoïsme* est plus proche de *logement* que de *santé*, contrairement à ce que pourrait laisser penser le plan factoriel de la figure 3.1.

Il ne faudrait pas en déduire que la classification donne des résultats plus précis que l'analyse des correspondances. Le bas de l'arbre (en fait, la partie gauche, pour les figures 4.1 à 4.4) donne effectivement des idées plus précises sur les distances locales ; mais, on l'a vu, les branches peuvent pivoter sur elles-mêmes, et la mise à plat de l'arbre ne donne que peu d'information sur les dispositions relatives des constellations plus importantes.

4.2.2 Classification des textes

La classification des colonnes de la table de contingence 3.1 du chapitre précédent se fait à partir des profils colonnes du tableau 3.3 que nous n'avons pas reproduit ici.

Le principe de l'agrégation est en tout point identique à ce qui vient d'être montré sur les lignes, et nous serons donc plus brefs dans les commentaires du tableau 4.3 et de la figure 4.5, homologues des tableaux 4.2 et de la figure 4.4.

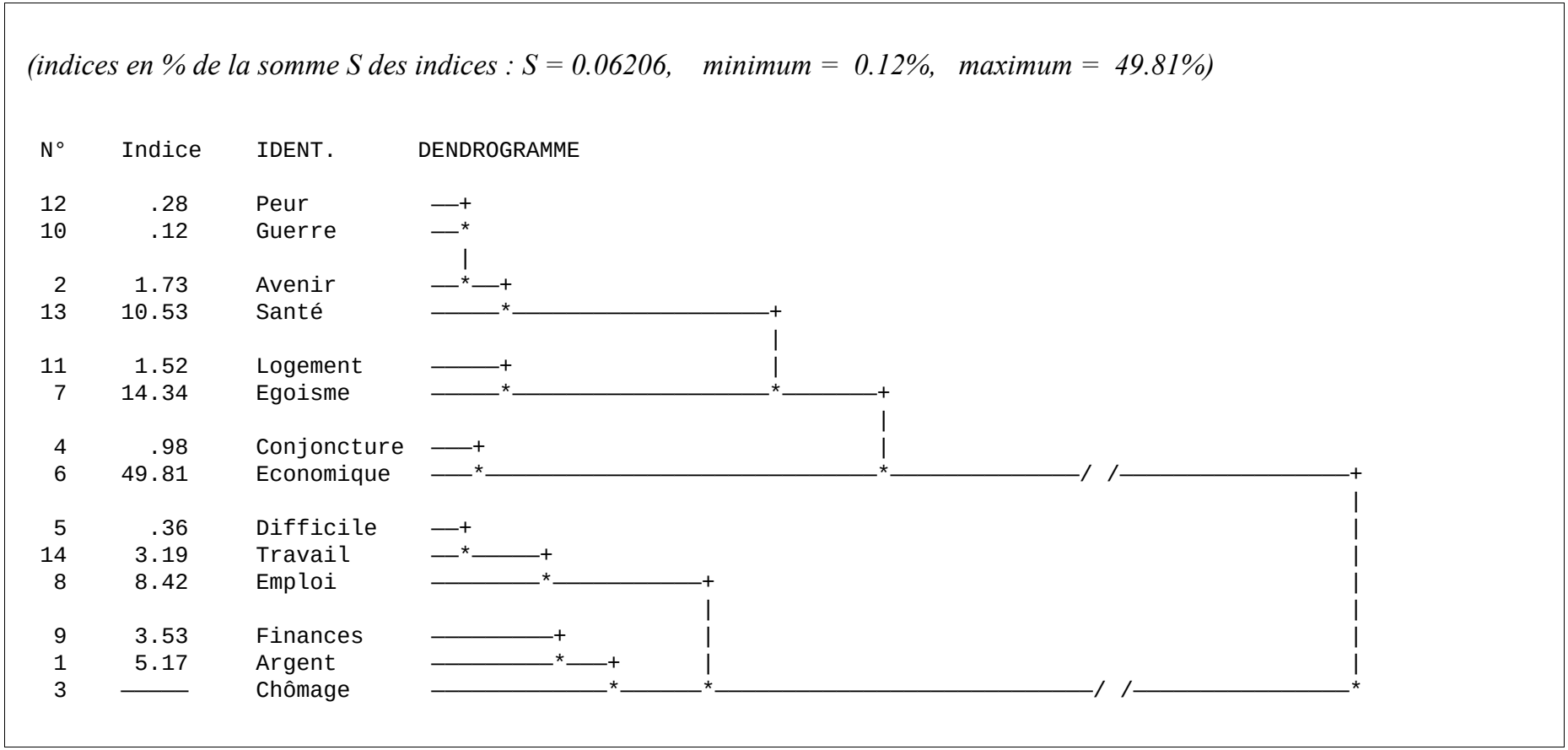

Figure 4.4

Dendrogramme décrivant les proximités entre lignes (formes) du tableau 4.1

Le tableau 4.3 décrit de façon similaire les étapes du fonctionnement de l'algorithme.

Tableau 4.3

Agrégation hiérarchique des colonnes : niveaux de diplômes.

NUM.	AINE	BENJ	EFF.	POIDS	INDICE	HISTOGRAMME DES INDICES DE NIVEAU
6	2	1	2	860.00	.00816	**********
7	4	3	2	607.00	.00950	**********
8	7	5	3	732.00	.01171	**********
9	8	6	5	1592.00	.03269	**********/ /*****

SOMME DES INDICES DE NIVEAU = .06206

Il y a 5 éléments à classer, et le premier noeud, qui porte le numéro 6, est formé des catégories 2 et 1 (numéros des colonnes sur le tableau 4.1) (*Sans diplôme* et *CEP*) qui totalisent 860 occurrences. Ce sont ensuite les catégories *Bacc* et *BEPC*, qui sont rejoint par la catégorie *Université*...

La figure 4.5 schématise le processus, en montrant comme précédemment que les deux branches principales de l'arbre correspondent à une opposition sur le premier axe factoriel de la figure 3.1 du chapitre précédent.

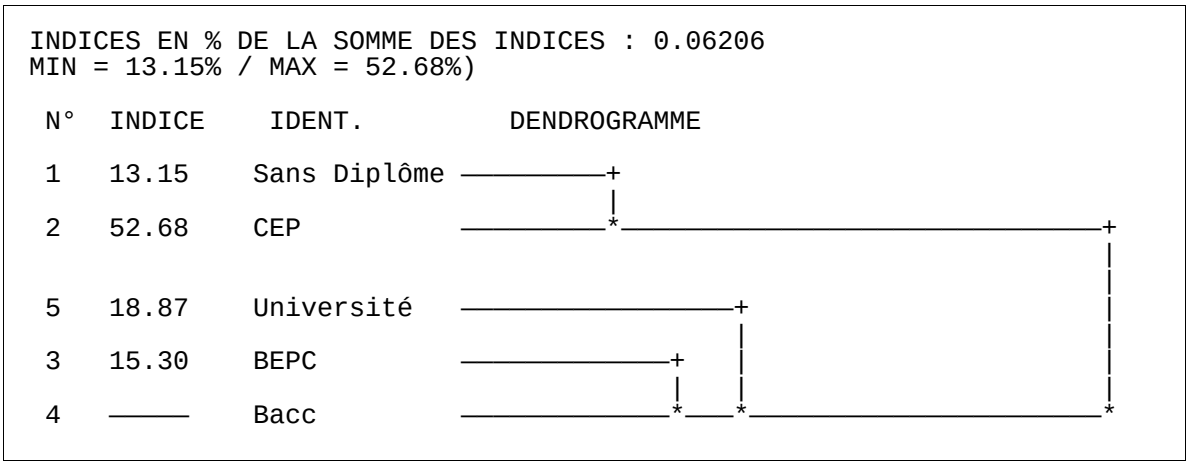

Figure 4.5

Dendrogramme décrivant les proximités entre colonnes du tableau 3.1

On remarque que la somme des indices a toujours la même valeur de 0.062, ce qui était attendu, car cette quantité fait intervenir de façon symétrique les lignes et les colonnes de la table. Toutefois, contrairement à ce qui se passe en analyse des correspondances, la classification des colonnes ne se déduit pas de façon simple de celle des lignes. Il n'existe pas d'équivalent des formules de transition.

4.2.3 Remarques sur les classifications de formes

Bien qu'une certaine symétrie semble relier sur le plan formel les classifications effectuées sur l'ensemble des parties et celles réalisées à partir de l'ensemble des formes, la pratique de la classification sur les tableaux lexicaux montre que les deux types d'analyse répondent à des besoins distincts qui entraînent, dans les deux cas, des utilisations différentes de la méthode.

Lorsqu'il s'agit d'étudier des textes (littéraires, politiques, historiques) les classifications portant sur les formes d'un corpus concernent en général des ensembles dont la dimension dépasse très largement celle de l'ensemble des parties. L'arbre de classification réalisé à partir d'un tel ensemble se présente sous une forme relativement volumineuse qui complique considérablement toute synthèse globale.

Dans la pratique, on abordera l'étude de la classification ainsi réalisée en considérant par priorité les associations qui se réalisent aux deux extrémités du dendrogramme :

a) les classes du niveau inférieur de la hiérarchie constituées par des agrégations de formes correspondant à un indice très bas (i.e. les classes agrégées dès le début de la classification)

b) les classes supérieures, souvent constituées de nombreuses formes, que l'on étudiera comme des entités. On retrouve en général au niveau des classes supérieures les principales oppositions observables dans les premiers plans factoriels.

Associations du niveau inférieur de la hiérarchie

Intéressons-nous tout d'abord aux associations réalisées aux tout premiers niveaux de la classification. Par construction, ces associations regroupent des ensembles de formes dont les profils de répartition sont très similaires (et parfois mêmes identiques) dans les parties du corpus.

Comme on va le voir, seul le retour systématique au contexte permet de distinguer parmi ces associations celles qui proviennent essentiellement de la reprise de segments plus ou moins longs, celles qui sont générées par les cooccurrences répétées de plusieurs formes à l'intérieur de mêmes phrases ou de mêmes paragraphes et celles des associations qui résultent de l'identité plus ou moins fortuite de la ventilation de certaines formes.

Un exemple de "quasi-segments"[1]

Dans un corpus de textes syndicaux, composé de neuf textes relatifs à quatre centrales syndicales[2], on a observé une identité parfaite de la ventilation des formes *pèse* et *lourdement*, identité qui a entraîné l'agrégation de ces deux formes dès le début du processus de classification.

Tableau 4.4

Distribution de deux formes dans neuf textes

	DT1	DT2	DT3	TC	GT1	GT2	GT3	FO1	FO2
lourdement	0	0	0	3	4	2	1	1	0
pèse	0	0	0	3	4	2	1	1	0

Cette identité laisse deviner la présence, dans différentes parties du corpus, des occurrences du stéréotype *"peser lourdement (sur quelque chose ou quelqu'un)*".

Le retour au contexte permet tout à la fois de vérifier l'existence de cette locution dans plusieurs des parties du corpus et de constater que les cooccurrences de ces deux formes ne se produisent pas exclusivement dans ce type de contexte.

L'examen des ventilations qui correspondent à ces deux formes montre tout d'abord qu'elles sont employées, dans ce corpus du moins, dans des contextes du type suivant :

> [CFTC] /.../ la prise en charge par les collectivités locales de l'entretien des espaces verts qui **pèse lourdement** sur les bénéficiaires des logements sociaux.
>
> [CGT1] une nouvelle orientation de la fiscalité /.../ est indispensable pour s'attaquer à l'inflation qui **pèse lourdement** sur les travailleurs et les titulaires de petits revenus et qui a également des effets négatifs sur les échanges extérieurs.

A côté de ces contextes, les deux formes apparaissent dans des contextes légèrement modifiés par rapport au segment considéré ci-dessus.

---

1 Cf. également sur ce point Becue (1993) à qui nous empruntons cette dénomination.

2 Ce corpus, réuni par J. Lefèvre et M. Tournier est composé de textes votés en congrès par les principales centrales syndicales françaises entre 1971 et 1976. Les parties sont notées : DT (CFDT), TC (CFTC), GT (CGT), FO (CGT-Force-Ouvrière). On trouvera des développements qui concernent ce problème dans Salem (1993).

> [CFTC] le comité national de la c-f-t-c-, /.../, constate que les prix de détail continuent à monter et que l'inflation **pèse** de plus en plus **lourdement** sur toutes les catégories sociales à revenu modeste, surtout les personnes âgées et les chargés de familles.

Enfin, chacune des deux formes peut apparaître dans des contextes libres des occurrences de l'autre forme : on retrouve parfois d'autres formes liées à cette même locution qui relativisent l'absence de la forme *lourdement.* Tel est le cas par exemple pour le contexte de la forme *pèse* :

> [CFTC] pour que l'europe **pèse** de **tout son poids** en face des deux grands blocs économiques

ou encore [CGT1] :

> l'économie américaine **pèse** d'un **poids déterminant.**

Cette identité de ventilation dans les parties du corpus des deux formes appelle un double commentaire. Tout d'abord, la ressemblance globale des profils résulte essentiellement de l'existence d'une expression récurrente, plus ou moins figée, qui contient les deux formes.

Cependant, l'identité stricte des deux ventilations est, dans le cas considéré, relativement fortuite, puisqu'elle résulte de l'addition des occurrences des deux formes liées à cette expression avec d'autres occurrences des deux formes présentes dans des contextes indépendants.

Les méthodes de la classification permettent de généraliser la recherche des cooccurrences de couples de formes à l'intérieur d'une même phrase au repérage de cooccurrences à l'intérieur de contextes plus étendu pouvant concerner plusieurs formes.

Classes de ventilations homogènes

On obtient une classification de l'ensemble des formes en un petit nombre de classes en procédant à une coupure de l'arbre correspondant à un niveau suffisamment élevé dans la hiérarchie indicée des classes.

Cette manière de procéder permet de considérer des classes qui regroupent, par construction, des formes dont la ventilation est relativement homogène. L'analyse du contenu de ces classes se fait en retournant fréquemment au contexte pour tenter de trouver les raisons profondes des similitudes que l'on constate entre les profils des formes qui appartiennent à une même classe.
Cette démarche permet également de vérifier que les occurrences des formes graphiques qui correspondent aux différentes flexions d'un même lemme se retrouvent affectées à une même classe, ce qui sera le signe d'une ventilation analogue dans les parties du corpus.

Classes périphériques

Toutes les classes découpées par l'algorithme de classification ne présentent pas le même intérêt. En projetant les centres de gravité de chacune des classes sur les premiers plans factoriels issus de l'analyse des correspondances[1], il est possible de repérer la correspondance de certaines des classes avec les types extrêmes dégagés par l'analyse des correspondances et par là même de sélectionner des ensembles de formes qui illustrent tout particulièrement ces grands types.

On considérera avec une attention toute particulière les groupes de formes qui s'associent très bas lors de la construction de l'arbre de classification et qui se rattachent relativement haut au reste des autres formes.

Documentation par les données segmentales

Ici encore, la lisibilité des classifications obtenues sur les formes s'accroît notablement si l'on prend soin de situer les ventilations relatives aux segments répétés parmi les résultats similaires obtenus sur les formes graphiques.

Comme dans le cas de l'analyse des correspondances, cette implication des segments peut se réaliser de deux manières différentes : analyse directe sur le tableau des formes et des segments répétés pour laquelle tous les éléments sont impliqués en qualité d'éléments principaux, ou implication des segments en qualité d'éléments supplémentaires (ou illustratifs) sur la base d'une classification préalable opérée sur les seules formes graphiques.

L'algorithme d'adjonction d'éléments supplémentaires à une classification (cf. Jambu, 1978) permet de rattacher un par un aux différentes classes de la hiérarchie indicée un ensemble d'individus statistiques n'ayant pas participé à la classification sur les éléments de base sans que cette dernière en soit aucunement affectée. Ici encore nous utiliserons massivement cette propriété pour documenter la classification par des calculs relatifs aux segments répétés. Comme on l'a annoncé plus haut, on peut également soumettre à la classification hiérarchique l'ensemble constitué par la réunion des termes (formes et segments répétés) dont la fréquence dépasse un certain seuil dans le corpus. Bien que cette seconde manière de procéder modifie les résultats

[1] On fait ici allusion aux techniques de calcul des contributions des classes aux facteurs d'une analyse des correspondances Jambu (1978).

obtenus sur les formes, l'expérience montre que les deux procédures ne conduisent pas, en règle générale, à des résultats trop éloignés.

Les nouvelles classes contiennent alors des formes et des segments dont la ventilation est relativement homogène dans les parties du corpus. L'étude de la disposition relative d'une forme et des segments qu'elle contient se trouve grandement facilitée par cette procédure.

Influence de la partition du corpus

On a vu plus haut que les distances mutuelles entre les différentes formes de l'ensemble soumis à la classification sont calculées à partir de leur ventilation dans les parties du corpus. Il s'ensuit que le découpage du corpus en parties revêt une importance primordiale pour la construction d'une classification à partir de l'ensemble des formes d'un même corpus car les variations dans la partition du corpus ont pour effet de rapprocher certains couples de formes et d'en éloigner d'autres ce qui influe forcément sur les classes de la hiérarchie.

Avec l'évolution croissante des capacités des ordinateurs, qui concerne tant les capacités mémoire que la vitesse d'exécution des calculs, les méthodes de la classification ascendante hiérarchique trouveront des applications dans le domaine de la recherche des cooccurrences textuelles à l'intérieur de phrases, de segments de texte de longueur fixe ou de paragraphes. Les algorithmes utilisés ici devraient permettre d'étendre les recherches que nous avons décrites plus haut en s'appuyant sur des partitions plus fines d'un même corpus correspondant à chacune de ces unités.

4.3 La Classification des fichiers d'enquête

Dans le cas du traitement statistique des fichiers d'enquêtes en vraie grandeur, les inconvénients théoriques et pratiques d'une démarche limitée aux méthodes de type factoriel sont particulièrement importants : ils concernent à la fois la nature même des résultats, et la gestion du volume de ces résultats. On insistera ici sur deux points :

a) Les visualisations sont souvent limitées à un petit nombre de dimensions, deux le plus souvent (premier plan factoriel), alors que la dimension réelle du phénomène étudié peut être bien supérieure (cette dimension est mesurée par le nombre d'axes significatifs[1]).

[1] On appelle ici *axe significatif,* au sens statistique du terme, un axe qui explique une part de dispersion ne pouvant être imputée au hasard. On peut déterminer le nombre d'axes significatifs en utilisant une procédure de simulation.

b) Ces visualisations peuvent inclure des centaines de points, et donner lieu à des graphiques chargés ou illisibles, à des listages de coordonnées encombrants, peu synthétiques.

Il faut donc à ce stade faire appel de nouveau aux capacités de gestion et de calcul de l'ordinateur pour compléter, alléger et clarifier la présentation des résultats. L'utilisation conjointe de méthodes de classification automatique adaptées à ce type de problème et des analyses de correspondances permet de remédier à ces lacunes.

Lorsqu'il y a trop de points sur un graphique, il paraît utile de procéder à des regroupements en familles homogènes. Mais les algorithmes utilisés pour ces regroupements fonctionnent de la même façon, que les points soient situés dans un espace à deux ou à dix dimensions.

Autrement dit, l'opération va présenter un double intérêt : allégement des sorties graphiques d'une part, prise en compte de la dimension réelle du nuage de points d'autre part.

Une fois les individus regroupés en classes, il est facile d'obtenir une description automatique de ces classes : on peut en effet, pour les variables numériques comme pour les variables nominales, calculer des statistiques d'écarts entre les valeurs internes à la classe et les valeurs globales ; on peut également convertir ces statistiques en *valeurs-test* et opérer un tri sur ces valeurs-test. On obtient finalement, pour chaque classe, les modalités et les variables les plus caractéristiques.

Après quelques mots sur les algorithmes de classification adaptés aux grands ensembles de données, on présentera un exemple d'application qui prolongera l'analyse des correspondances multiples présentée au chapitre précédent (paragraphe 4.3), et qui illustrera cette nouvelle présentation compacte de l'information.

4.3.1 Les algorithmes de classification mixte

L'algorithme de classification qui nous paraît le plus adapté au partitionnement d'un ensemble comprenant des milliers d'individus est un *algorithme mixte* procédant en trois phases :

a) *Partitionnement initial* en quelques dizaines de classes par une technique du type "nuées dynamiques" ou "k-means", Diday (1971).
On peut résumer sans trop les trahir ces techniques en quelques phrases. On commence par tirer au hasard des individus qui seront des centres provisoires de classes. Puis, on affecte tous les individus au centre provisoire le plus proche (au sens d'une distance telle que la distance du

chi-2 définie au chapitre précédent). On construit ainsi une partition de l'ensemble des individus. On calcule de nouveau des centres provisoires, qui sont maintenant les "centres" (points moyens par exemple) des classes qui viennent d'être obtenues, et on réitère le processus, autrement dit on affecte de nouveau tous les individus à ces centres, ce qui induit une nouvelle partition, etc. Le processus se stabilise nécessairement, mais la partition obtenue dépendra en général du choix initial des centres.[1]

b) *Agrégation hiérarchique des classes obtenues* : Les techniques de classification hiérarchique présentées dans ce chapitre sont assez coûteuses si elles s'appliquent à des milliers d'éléments, c'est pourquoi il faut réduire la dimension du problème en opérant un regroupement préalable en quelques dizaines de classes.

L'intérêt du dendrogramme obtenu est qu'il peut donner une idée du nombre de classes existant effectivement dans la population. Chaque coupure d'un tel arbre va donner une partition, ayant d'autant moins de classes que l'on coupe près du sommet.

c) *Coupure de l'arbre* (en général après une inspection visuelle)

Plus on agrège de points, autrement dit plus on se rapproche du sommet de l'arbre, plus la distance entre les deux classes les plus proches est grande. On a vu (paragraphe 4.1.2) qu'en coupant l'arbre au niveau d'un saut important de l'indice, on peut espérer obtenir une partition de bonne qualité, car les individus regroupés auparavant étaient proches, et ceux regroupés après la coupure seront nécessairement éloignés, ce qui est la définition d'une bonne partition.

d) *Optimisation de la partition obtenue par réaffectations*

Une partition obtenue par coupure n'est pas la meilleure possible, car l'algorithme de classification hiérarchique n'a malheureusement pas la propriété de donner à chaque étape une partition optimisée. On peut encore améliorer la partition obtenue par réaffectation des individus comme indiqué au paragraphe a) ci-dessus. Malgré la relative complexité de la procédure, on ne peut être assuré d'avoir trouvé la "meilleure partition en k classes".

[1] La recherche d'une partition optimale vis-à-vis d'un critère tels que ceux que l'on utilise habituellement en statistique (par exemple maximisation du quotient variance externe / variance interne) se heurte encore à l'*infini combinatoire* et sa mise en évidence n'est actuellement pas possible même sur les ordinateurs les plus puissants.

4.3.2 Séquence des opérations en traitements d'enquêtes

Les analyses complètes incluant la phase de filtrage par analyse des correspondances se dérouleront en définitive de la façon suivante :

a) Choix des éléments actifs (qui correspond au choix d'un point de vue). On peut décrire les individus du point de vue de leurs caractéristiques de base, mais aussi à partir d'un thème particulier : habitudes de consommation, opinions politiques, etc.

b) Analyse des correspondances simples ou multiples à partir de ces éléments actifs.

c) Positionnement des éléments illustratifs. On projettera ainsi toute l'information disponible de nature à comprendre ou à interpréter la typologie induite par les éléments actifs.

d) Examen des premiers graphiques de plans factoriels en général limités aux points occupant les positions les plus significatives.

e) Partition de l'ensemble des individus ou observations selon la méthode qui vient d'être exposée.

f) Position sur les graphiques précédents des centres des principales classes (une partition définit toujours une variable nominale particulière).

g) Description des classes par les modalités et les variables les plus caractéristiques.

4.3.3 Exemple d'application : Situations-types ou noyaux factuels

Cet exemple prolonge l'analyse des correspondances multiples du chapitre précédent. Les 144 individus enquêtés sont décrits par le même ensemble de variables actives. On veut maintenant obtenir un petit nombre de groupes d'individus les plus homogènes possible vis-à-vis de leurs caractéristiques de base. L'intérêt de tels regroupements apparaîtra au chapitre 5, lorsqu'il s'agira d'agréger les réponses libres sans privilégier de critères particuliers.

On aimerait pouvoir croiser des caractéristiques telles que l'âge, le sexe, la profession, le niveau d'instruction, de façon à étudier des groupes d'individus tout à fait comparables entre eux du point de vue de leur situation objective, c'est-à-dire de réaliser, dans les limites du possible, le *toutes choses égales par ailleurs,* situation idéale, mais hors de portée. En pratique, de tels croisements conduisent vite à des milliers de modalités, dont on ne sait que faire lorsque l'on étudie un échantillon de l'ordre de 1 000 individus. De plus, les croisements ne tiennent pas compte du réseau d'interrelations existant entre

ces caractéristiques : certaines sont évidentes a priori (il n'y a pas de jeunes retraités), d'autres sont également connues a priori, mais peuvent souffrir des exceptions (il y a peu d'étudiants veufs, d'ouvriers diplômés d'université), d'autres enfin ont un caractère plus statistique (il y a plus de femmes dans les catégories employés et veufs).

En pratique, on peut espérer trouver des regroupements opératoires en une vingtaine de classes pour un échantillon de l'ordre de 2 000 individus. On verra qu'un regroupement plus sommaire en cinq grandes classes n'est pas totalement dépourvu d'intérêt, en restant compatible avec le volume restreint d'un exemple pédagogique. De telles classes sont appelées *situations-types* ou encore *noyaux factuels*. Elles seront utilisées au chapitre 5 pour procéder à des regroupements de réponses libres (paragraphe 5.2).

Application à l'exemple du paragraphe 3.3.1 (chapitre 3)

La procédure de classification mixte décrite ci-dessus a été appliquée aux 144 individus partir des cinq variables actives décrites au paragraphe 3.3.1. Quatre classes ont été obtenues par coupure du dendrogramme, puis optimisation. On présentera seulement ici la phase nouvelle de description automatique des classes.

Le tableau 4.5 va en effet décrire les classes de façon précise, en comparant les pourcentages de réponses internes aux classes aux pourcentages globaux, puis en sélectionnant les modalités les plus caractéristiques (cf. guide de lecture du tableau 4.5). On note que parmi les modalités les plus caractéristiques d'une classe figurent également des modalités illustratives, qui n'ont pas participé à la fabrication des classes. Tels sont les cas des modalités décrivant des appartenances professionnelles.

On voit assez bien les avantages que présentent les partitions pour décrire des ensembles multidimensionnels :

La notion de classe est intuitive (groupes d'individus les plus semblables possible). La description des classes fait appel à des classements de libellés complets et donc facile à lire, ces classements étant fondés sur de simples comparaisons de pourcentages. Il est donc plus facile de décrire des classes qu'un espace continu.

Tableau 4.5

Exemple de description automatique de la partition en 4 noyaux factuels de l'exemple du paragraphe 3.3.1 (chapitre 3)

MODALITES CARACTERISTIQUES		POURCENTAGES CLA/MOD	POURCENTAGES MOD/CLA	POURCENTAGES GLOBAL	POIDS	V.TEST	PROBA
- CLASSE 1 / 4				45.83	66		
Niveau de diplôme	*Baccalauréat*	93.02	60.61	29.86	43	7.61	.000
Niveau de diplôme	*BEPC*	72.22	39.39	25.00	36	3.50	.000
- CLASSE 2 / 4				13.89	20		
Niveau de diplôme	*Sans diplôme ou CEP*	74.07	100.00	18.75	27	8.82	.000
Profession	*EMPLOYE*	42.86	30.00	9.72	14	2.58	.005
Age en 3 classes	*Entre 30 et 50 ans*	24.49	60.00	34.03	49	2.33	.010
- CLASSE 3 / 4				26.39	38		
Niveau de diplôme	*Université*	100.00	100.00	26.39	38	12.42	.000
Age en 3 classes	*Moins de 30 ans*	36.49	71.05	51.39	74	2.66	.004
Profession	*CADRESUP, PROF LIB*	75.00	15.79	5.56	8	2.62	.004
Profession	*ETUDIANT*	40.43	50.00	32.64	47	2.42	.008
- CLASSE 4 / 4				13.89	20		
Age en 3 classes	*Plus de 50 ans*	95.24	100.00	14.58	21	9.94	.000
Avez-vous eu des enfants	*oui*	28.85	75.00	36.11	52	3.59	.000
Etes vous actuellement	*Veuf(ve)*	100.00	20.00	2.78	4	3.45	.000
Sexe	*femme*	21.62	80.00	51.39	74	2.57	.005
Profession	*MENAGERE*	57.14	20.00	4.86	7	2.44	.007
Niveau de diplôme	*BEPC*	27.78	50.00	25.00	36	2.39	.008
Profession	*ARTISANTS, COMMERC.*	75.00	15.00	2.78	4	2.39	.009

Mais l'analyse des correspondances permet de visualiser les positions relatives des classes dans l'espace, et aussi de mettre en évidence certaines variations continues ou dérives dans cet espace qui auraient pu être masquées par la discontinuité des classes. Les deux techniques sont donc complémentaires, et se valident mutuellement.

Guide de lecture du tableau 4.5

Nous commencerons par les deux dernières colonnes, qui sont en fait les plus importantes puisqu'elles permettent de sélectionner et de classer les modalités caractéristiques de chaque classe.

Pour chacune des classes, les modalités les plus caractéristiques sont rangées suivant les valeurs décroissantes de la valeur-test V.TEST (avant dernière colonne du tableau), ou, ce qui revient au même, selon les valeurs croissantes de la probabilité PROBA (dernière colonne).

La valeur-test V.TEST est, brièvement, l'analogue de la valeur absolue d'une variable normale centrée réduite, qui est significative au seuil 5% bilatéral si elle dépasse la valeur 1.96, ce qui est le cas pour toutes les modalités retenues ici.

La première colonne numérique CLA/MOD donne les pourcentages de chaque classe dans les modalités : ainsi, tous les veufs (ves) de l'échantillon appartiennent à la classe 4 (CLA/MOD = 100).

La seconde colonne numérique donne le pourcentage interne MOD/CLA de chaque modalité dans la classe. Ainsi, il y a 20% de veufs(ves) dans la classe 4 (MOD/CLA = 20). Ici, ce pourcentage interne est nécessairement plus élevé que le pourcentage global (troisième colonne : GLOBAL) puisque les modalités sélectionnées sont caractéristiques des classes. C'est précisément la différence entre le pourcentage interne et le pourcentage global qui est à la base du calcul de la valeur-test.

Ainsi, la classe 3 contient 38 individus, soit 26% de la population. La modalité "Cadre Sup.", qui concerne 5.56 % de l'échantillon, concerne près de 15.79% des personnes de cette classe. Avec une valeur-test de 2.62, elle vient en quatrième rang pour caractériser la classe.

La troisième colonne, GLOBAL, donne, on l'a vu, le pourcentage de chaque modalité dans la population globale. Il s'obtient simplement en divisant la colonne poids par 144, effectif global de l'échantillon.

La quatrième colonne, POIDS, donne les effectifs bruts des classes et des modalités.