

## 3 Analyse Factorielle des Correspondances

### 3.1 Introduction

L'analyse factorielle des correspondances (AFC), ou analyse des correspondances simples, est une méthode exploratoire d'analyse des tableaux de contingence. Elle a été développée essentiellement par J.-P. Benzecri durant la période 1970-1990.

Soient deux variables nominales X et Y, comportant respectivement p et q modalités. On a observé les valeurs de ces variables sur une population et on dispose d'un tableau de contingence à p lignes et q colonnes donnant les effectifs conjoints c'est-à-dire les effectifs observés pour chaque combinaison d'une modalité i de X et d'une modalité j de Y.

Les valeurs de ce tableau seront notées  $n_{ij}$ , l'effectif total sera noté N.

L'ACP vise à analyser ce tableau en apportant des réponses à des questions telles que :

- Y a-t-il des lignes du tableau (modalités de X) qui se "ressemblent", c'est-à-dire telles que les distributions des modalités de Y soient analogues ?
- Y a-t-il des lignes du tableau (modalités de X) qui s'opposent, c'est-à-dire telles que les distributions des modalités de Y soient très différentes ?
- Mêmes questions pour les colonnes du tableau.
- Y a-t-il des associations modalité de X - modalité de Y qui s'attirent (effectif conjoint particulièrement élevé) ou qui se repoussent (effectif conjoint particulièrement faible) ?

La méthode se fixe également comme but de construire des représentations graphiques mettant en évidence ces propriétés des données.

### 3.2 Analyse factorielle des correspondances avec Statistica

#### 3.2.1 Traitement des données avec Statistica

Source : Exemple fourni avec le logiciel Statistica.

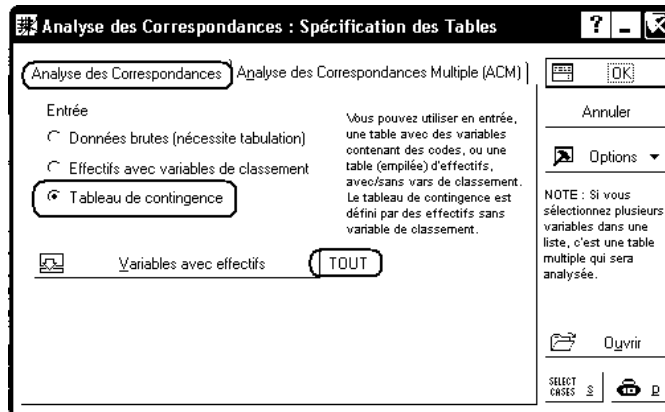
Supposons que vous ayez collecté des données sur les habitudes de différents salariés d'une entreprise concernant la cigarette. Les données suivantes sont présentées dans l'ouvrage de Greenacre (1984, p. 55).

Ouvrez le classeur Smoking.stw et observez les 3 feuilles de données saisies.

Commençons, par exemple, par rendre active la feuille de données Smoking1.sta (tableau de contingence).

	Analyse des correspondances simple.			
	1 NON_FUM	2 OCCAS	3 MOYEN	4 GROS_FUM
CADRE_EXPER	4	2	3	2
CADRE_DEBUT	4	3	7	4
EMPLOY_EXPER	25	10	12	4
EMPLOY_DEBUT	18	24	33	13
SECRETAIRES	10	6	7	2

Pour effectuer l'AFC, nous utilisons le menu Statistiques - Techniques exploratoires multivariées - Analyse des correspondances.



La fenêtre de dialogue permet d'indiquer la manière dont se présentent nos données. La situation la plus classique est celle d'un tableau de contingence : les modalités lignes sont indiquées dans une variable spécifiques, les modalités colonnes sont les autres variables du tableau, et la feuille de données contient les effectifs  $n_{ij}$ .

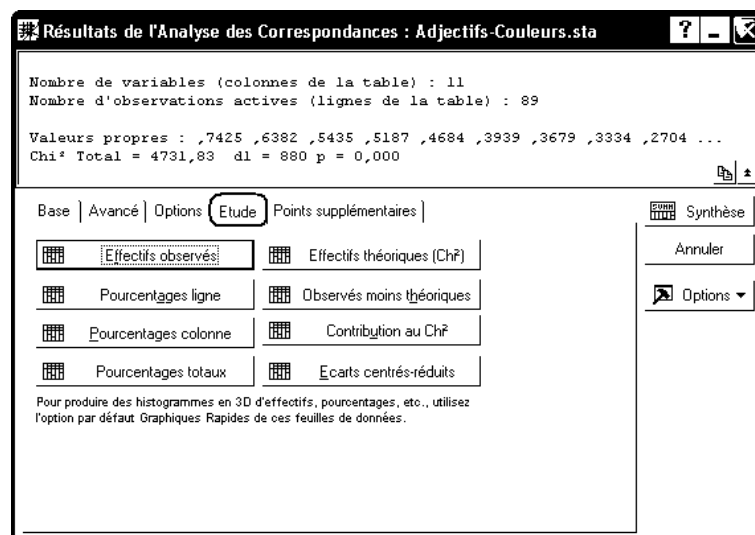
On indique également les variables qui participeront à l'analyse (ici toutes les variables). Notez que les zéros éventuels sont obligatoires, car une cellule laissée vide est interprétée comme une valeur manquante, et c'est alors l'ensemble de la ligne qui est éliminé de l'analyse.

N.B. Ne fermez pas l'analyse en cours pendant la suite des manipulations. Ainsi, vous n'aurez pas à indiquer de nouveau les options ci-dessus, vos résultats seront cohérents entre eux et se rassembleront dans un même classeur.

### 3.2.1.1 Statistiques descriptives

Les principaux résultats de statistiques descriptives pourront être obtenus à partir de l'onglet "Etude".

On peut ainsi obtenir les fréquences, les fréquences lignes, les fréquences colonnes et les profils moyens.



Par exemple, le tableau des fréquences et les profils ligne et colonne moyens sont :

Pourcentages Totaux (Smoking1.sta dans Smoking.stw)  
 Inertie Totale = ,08519 Chi² = 16,442 dl = 12 p = ,17190

	NON_FUM	OCCAS	MOYEN	GROS_FUM	Total
CADRE_EXPER	2,07	1,04	1,55	1,04	5,70
CADRE_DEBUT	2,07	1,55	3,63	2,07	9,33
EMPLOY_EXPER	12,95	5,18	6,22	2,07	26,42

EMPLOY_DEBUT	9,33	12,44	17,10	6,74	45,60
SECRETAIRES	5,18	3,11	3,63	1,04	12,95
Total	31,61	23,32	32,12	12,95	100,00

Remarques :

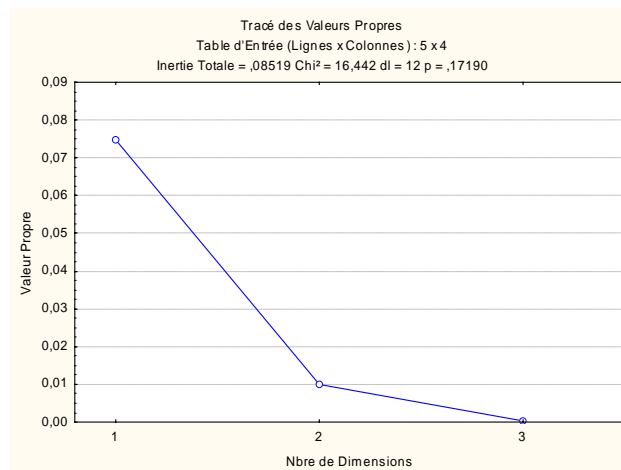
1) Dans cet exemple, le niveau de significativité du khi-2 n'est que de 17%. Autrement dit, la position dans l'entreprise et le comportement vis-à-vis du tabac sont, dans une large mesure, indépendantes. Mais le but de l'AFC est de mettre en évidence les "ressemblances" ou les "dissemblances" entre lignes ou entre colonnes, et la méthode fonctionne même si les différences sont de faible amplitude.

2) Statistica ne permet pas d'obtenir directement le tableau des taux de liaison, qui est pourtant un outil exploratoire intéressant. Mais on pourra utiliser les tableaux "Observés moins théoriques" et "Contribution au  $\chi^2$ ".

### 3.2.1.2 Choix des valeurs propres

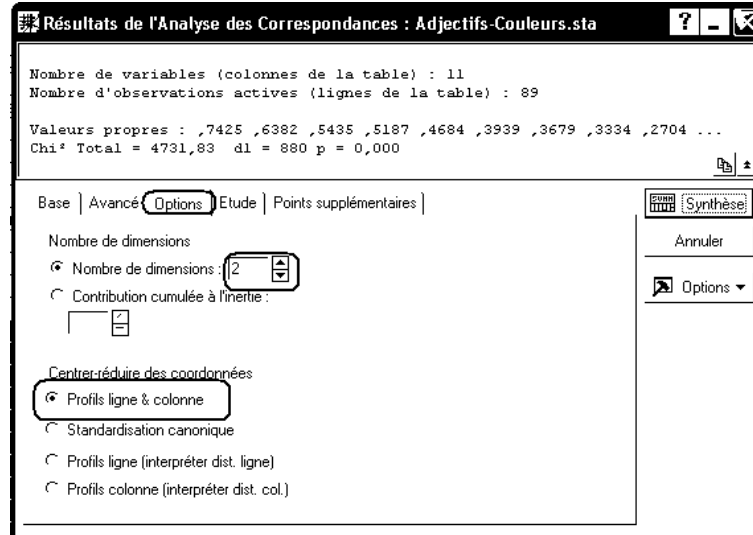
C'est ensuite l'onglet "Avancé" qui nous permettra d'afficher les valeurs propres, et donc de choisir le nombre d'axes à garder.

Valeurs Propres et Inertie de toutes les Dimensions (Smoking1.sta dans Smoking.stw)					
Table d'Entrée (Lignes x Colonnes) : 5 x 4					
Inertie Totale = ,08519 $\chi^2$ = 16,442 dl = 12 p = ,17190					
Nombre de Dims.	ValSing.	ValProp.	%age Inertie	%age Cumulé	$\chi^2$
1	0,2734	0,0748	87,76	87,76	14,43
2	0,1001	0,0100	11,76	99,51	1,93
3	0,0203	0,0004	0,49	100,00	0,08



### 3.2.1.3 Résultats relatifs aux individus-lignes et aux individus-colonnes.

Pour les résultats qui suivent, on indique le nombre d'axes factoriels à conserver sous l'onglet "Base" ou sous l'onglet "Options". Ce dernier permet également de choisir plusieurs types d'échelles pour représenter lignes et colonnes. Le type de représentation le plus classique, qui fait jouer des rôles symétriques aux lignes et aux colonnes, correspond à la première option.

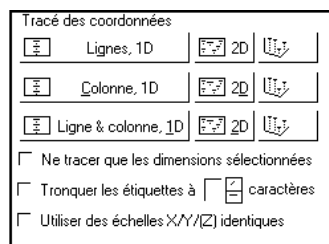


On retourne ensuite sous l'onglet "Avancé" pour afficher les coordonnées des individus-lignes et des individus-colonnes. On notera que Statistica produit deux tableaux de résultats, et on passera de l'un à l'autre à l'aide des onglets du classeur.

Coordonnées Colonne et Contributions à l'Inertie (Smoking1.sta dans Smoking.stw)										
Table d'Entrée (Lignes x Colonnes) : 5 x 4										
Standardisation : Profils ligne et colonne										
Nom Col.	Colonne Numéro	Coord. Dim.1	Coord. Dim.2	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus <sup>2</sup> Dim.1	Inertie Dim.2	Cosinus <sup>2</sup> Dim.2
NON_FUM	1	-0,3933	0,0305	0,3161	1,0000	0,5774	0,6540	0,9940	0,0293	0,0060
OCCAS	2	0,0995	-0,1411	0,2332	0,9840	0,0829	0,0308	0,3267	0,4632	0,6573
MOYEN	3	0,1963	-0,0074	0,3212	0,9832	0,1480	0,1656	0,9818	0,0017	0,0014
GROS_FUM	4	0,2938	0,1978	0,1295	0,9946	0,1917	0,1495	0,6844	0,5058	0,3102

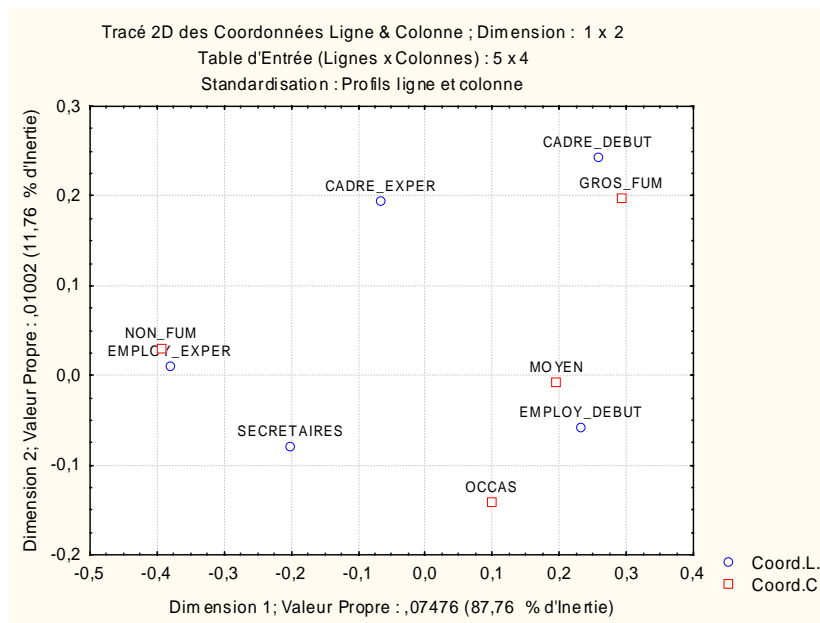
Coordonnées Ligne et Contributions à l'Inertie (Smoking1.sta dans Smoking.stw)										
Table d'Entrée (Lignes x Colonnes) : 5 x 4										
Standardisation : Profils ligne et colonne										
NomLigne	Ligne Numéro	Coord. Dim.1	Coord. Dim.2	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus <sup>2</sup> Dim.1	Inertie Dim.2	Cosinus <sup>2</sup> Dim.2
CADRE_EXPER	1	-0,0658	0,1937	0,0570	0,8926	0,0314	0,0033	0,0922	0,2136	0,8003
CADRE_DEBUT	2	0,2590	0,2433	0,0933	0,9911	0,1395	0,0837	0,5264	0,5512	0,4647
EMPLOY_EXPER	3	-0,3806	0,0107	0,2642	0,9998	0,4497	0,5120	0,9990	0,0030	0,0008
EMPLOY_DEBUT	4	0,2330	-0,0577	0,4560	0,9998	0,3084	0,3310	0,9419	0,1518	0,0579
SECRETAIRES	5	-0,2011	-0,0789	0,1295	0,9986	0,0711	0,0701	0,8653	0,0805	0,1333

On utilise ensuite les boutons du bloc "Tracé des coordonnées" pour obtenir des représentations graphiques des résultats de l'AFC.



Les graphiques "par axe" pourront être obtenus à l'aide du bouton "Ligne & colonne, 1D". Le graphique dans un plan, superposant les résultats des lignes et des colonnes, pourra être obtenu à l'aide du bouton "2D" de la même ligne. En revanche, il n'est pas évident d'éliminer certaines

étiquettes pour améliorer la lisibilité du graphique. La seule méthode paraît être de faire un clic droit sur une étiquette, de sélectionner l'item de menu "Propriétés..." puis d'éditer manuellement le tableau des étiquettes qui s'affiche.



### 3.2.1.4 Structures possibles pour les données d'entrée

Jusqu'à présent, nous avons travaillé à partir d'un tableau de contingence. Mais Statistica nous permet également de réaliser l'AFC à partir d'un tableau d'effectifs (feuille de données Smoking2.sta) ou d'un tableau protocole (données non recensées - feuille de données Smoking3.sta).

Refaites l'AFC précédente, d'abord en utilisant Smoking2.sta comme feuille de données active, puis en utilisant Smoking3.sta.

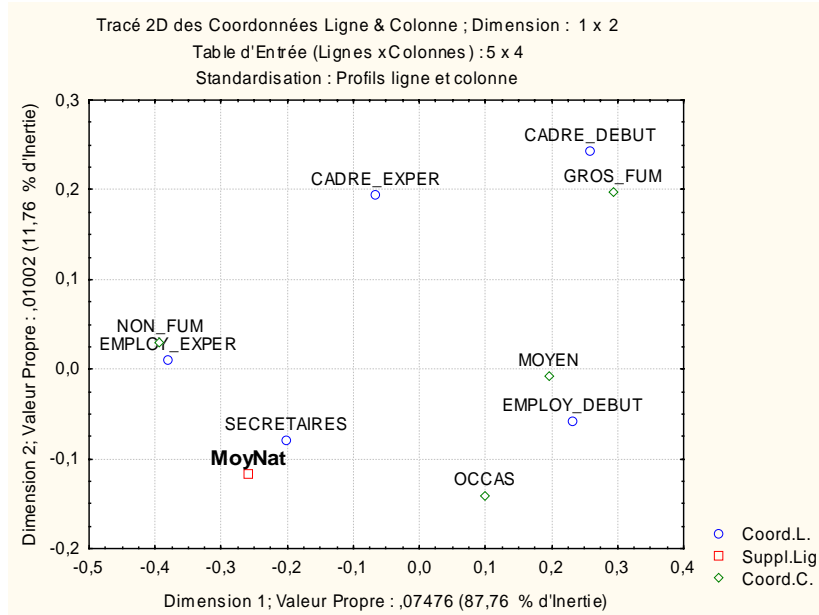
### 3.2.1.5 Ajout de lignes supplémentaires

On veut par exemple faire figurer sur le graphique un point correspondant à la distribution de la consommation de tabac dans la population nationale : non fumeurs : 42%, fumeurs occasionnels : 29%, fumeurs moyens : 20% et gros fumeurs : 9%. Utilisez pour cela l'onglet "Points supplémentaires", puis le bouton "Ajouter des points ligne".

Points Ligne Supplémentaires (Smoking1.sta dans Smoking.stw)

Points Ligne Supplémentaires (Smoking1.sta dans Smoking.stw)  
 Saisissez les valeurs (effectifs) des nouveaux points supplémentaires puis cliquez sur OK.

Point	Nom du Pt Suppl	NON_FUM	OCCAS	MOYEN	GROS_FUM
1	Moyenne Nationale	42	29	20	9
2					

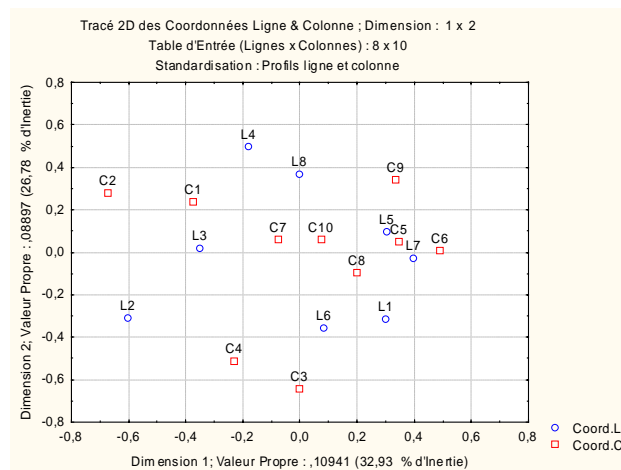


### 3.3 Quelques configurations remarquables dans les résultats produits par une AFC.

On pourra consulter le fichier Configurations-Types.stw qui rassemble quelques configurations classiques de nuages, générées à partir de données fictives.

#### 3.3.1 Forme générale du nuage

L'inertie totale (le Phi-2) est un indicateur de la dispersion totale du nuage. La comparaison des inerties de chacun des axes (c'est-à-dire des valeurs propres associées aux axes) renseigne sur la forme du nuage de points. Si les premières valeurs propres sont proches les unes des autres, la dispersion est relativement homogène : il n'y a pas vraiment de direction privilégiée et le nuage de points est approximativement sphérique. Si au contraire, les valeurs propres sont nettement différentes, cela traduit un nuage de points fortement allongé selon une (ou plusieurs) direction.

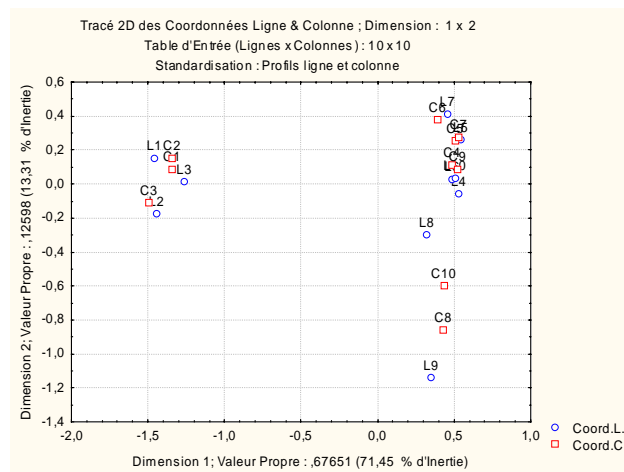


#### 3.3.2 Deux paquets de points - Valeurs propres proches de 1

Les valeurs propres sont toutes inférieures à 1. Mais, une valeur propre proche de 1 indique une dichotomie des données, c'est-à-dire un tableau de contingence qui, après reclassement des modalités, aurait l'allure suivante :

	0
0	

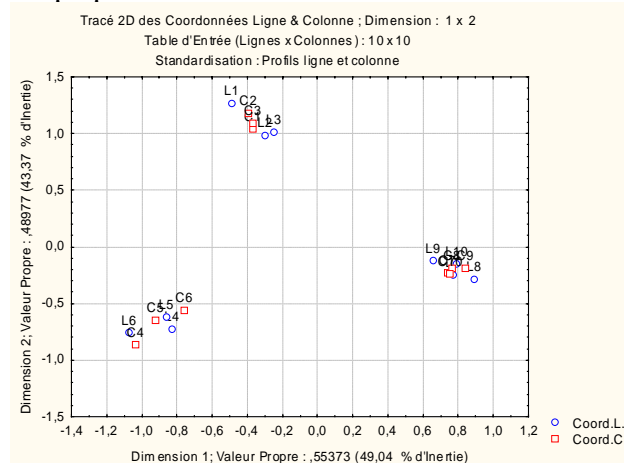
Le nuage est alors divisé en deux paquets de points. La feuille de données "Deux-paquets" fournit une illustration de cette situation.



### 3.3.3 Trois paquets de points

De même, l'existence de deux valeurs propres proches de 1 indique une partition des observations en 3 groupes. Si toutes les valeurs propres sont proches de 1, cela indique une correspondance entre chaque modalité ligne et une modalité colonne "associée". Avec une réorganisation convenable des modalités, les effectifs importants se trouvent alors le long de la diagonale.

La feuille de données "Trois-paquets" fournit une illustration de cette situation.

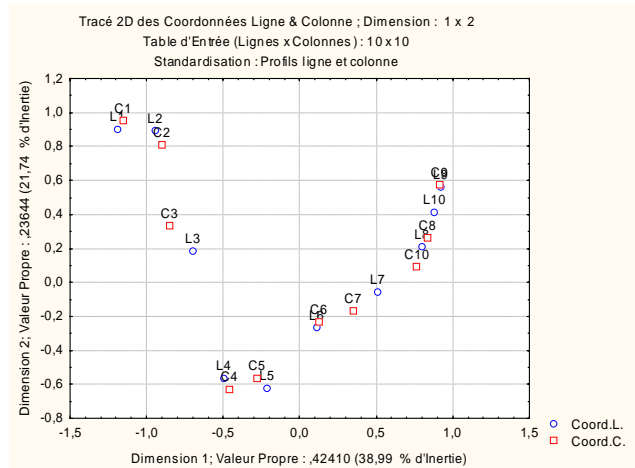
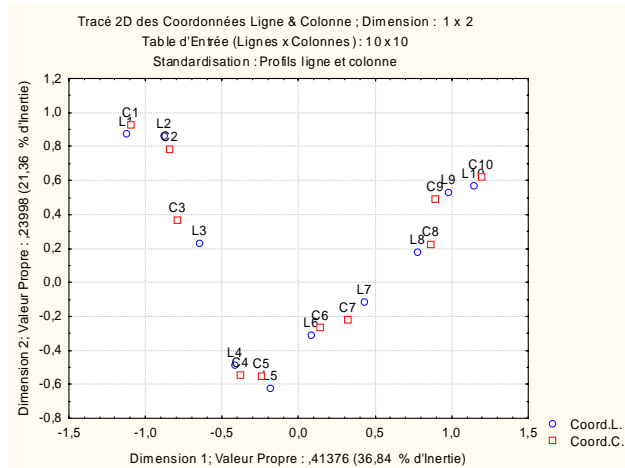


### 3.3.4 L'effet Guttman.

Un nuage de points de forme parabolique indique une redondance entre les deux variables étudiées : la connaissance de la ligne  $i$  donne pratiquement celle de la colonne  $j$ . Dans un tel cas, pratiquement toute l'information est contenue dans le premier facteur. Cette configuration se rencontre notamment lorsque les deux variables sont ordinales, et classent les sujets de la même façon. Dans ce cas, le premier axe oppose les valeurs extrêmes et classe les valeurs, tandis que le deuxième axe oppose les intermédiaires aux extrêmes.

La feuille de données "Effet-Guttman" fournit une illustration assez caractéristique de cette situation. Dans ce cas, on a intérêt à ne pas limiter l'étude au plan (1, 2). La configuration-type dans les trois plans de projection définis par les 3 premiers axes prend souvent les allures indiquées dans l'exemple.

Il pourra alors être intéressant d'examiner les accidents des courbes qui joignent les points, qui reflètent les particularités des situations étudiées. Voir par exemple la situation des modalités L10 et C10 dans l'exemple "Guttman-perturbé".



### 3.3.5 Nuage tétraédrique

Le premier exemple ("Deux-paquets") est également caractéristique d'une forme classique de nuage : tétraédrique, ou en forme de "berlingot" comme on peut s'en rendre compte en construisant les projections du nuage sur les 3 premiers axes.

### 3.4 L'extension de la notion de tableau de contingence

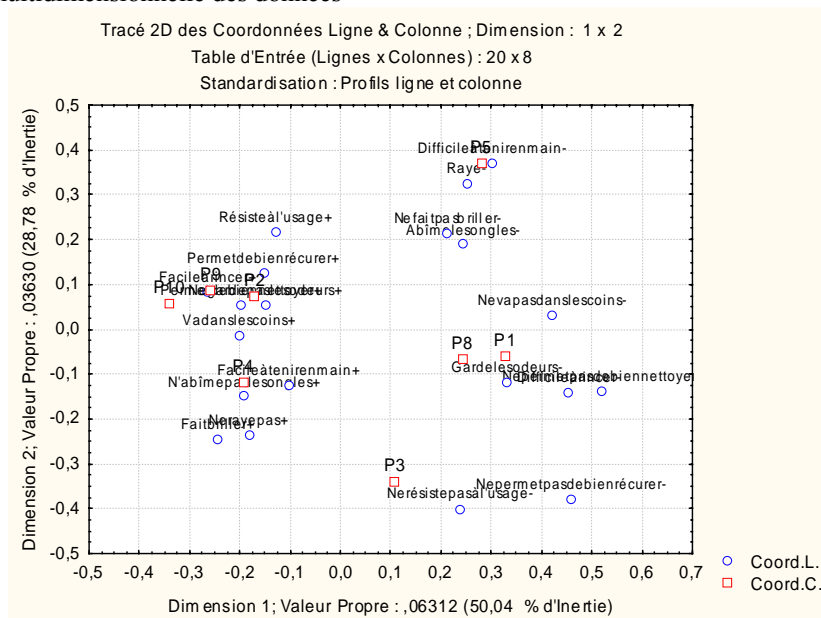
En toute rigueur, l'analyse de correspondances ne s'applique qu'aux tableaux de contingence. Elle peut cependant être appliquée à des tableaux qui, a priori, ne sont pas des tableaux de contingence. Un critère essentiel pour décider si un tableau peut être assimilé à un tableau de contingence est le suivant : on doit pouvoir donner un sens à la somme des cases du tableau, qu'elle soit faite par ligne ou par colonne.

#### 3.4.1 Exemple 1 :

Traiter par AFC les données du fichier Protein.sta.

#### 3.4.2 Tableaux juxtaposés

Considérons l'exemple fourni dans le classeur Echelles-Likert.stw. On obtient ainsi un point par produit et deux points par échelle bipolaire. On peut facilement montrer que le barycentre (pondéré) des deux points correspondant à une échelle donnée se trouve au centre de gravité du nuage. Si le point "+" se trouve plus près de l'origine que le point "-", cela signifie que l'intensité de la propriété positive est supérieure à celle de la propriété négative correspondante. Cet effet est connu sous le nom d'effet de levier.



Dans certains cas, on peut juxtaposer, par exemple, deux tableaux de contingence correspondant à des dates différentes, par exemple la ventilation de la population française par région et par CSP pour deux recensements différents. Il sera alors pertinent d'étudier comment chaque modalité s'est déplacée entre l'époque 1 et l'époque 2.

### 3.4.3 Juxtaposer plusieurs tableaux : vers l'ACM

Source : Hahn A., Eirmbter W. H., Jacob R., Le sida : savoir ordinaire et insécurité, traduction française de Herrmann M.

Il s'agit d'une enquête réalisée durant l'été 1990, auprès d'un échantillon représentatif des ménages de RFA.

Résumé du questionnaire :

Variable	Modalité	Codage
Sexe	masculin	m
	féminin	f
Confession	protestant	ev
	catholique	rk
	autre	an
	sans	ke
Liens avec l'église	forts	f1
	moyens	f2
	inexistants	f3
Catégorie Sociale	élèves/étud	s1
	classe sup.	s2
	cl. moy. sup.	s3
	cl. moyenne	s4
	cl. moy. inf.	s5
	cl. populaire	s6
	autres	s7

Taille du lieu de résidence	< 2	k1
	2 à < 5	k2
	5 à < 20	k3
	20 à < 50	k4
	50 à < 100	k5
	100 è < 500	k6
	> 500	k7
Classe d'âge	18 à < 30	a1
	30 à < 40	a2
	40 à < 50	a3
	50 à < 60	a4
	60 et plus	a5
Fidélité dans les rapports sexuels	très pour	t1
	plutôt pour	t2
	indécis	t3
	plutôt contre	t4
	très contre	t5
Plusieurs partenaires	oui	p1
	non	p2
Préférences politiques	CDU/CSU	cd
	SPD	sp
	FDP	fd
	Verts	gr
Nombre de situations jugées contaminantes	0	w0
	1	w1
	2	w2
	3	w3
	4	w4
	5	w5
	6	w6
	7	w7
	8	w8
Le sida est la conséquence d'une faute et d'une punition	très pour	c1
	plutôt pour	c2
	indécis	c3
	plutôt contre	c4
	très contre	c5
Dispositions d'évitement et d'exclusion des contaminés de la sphère personnelle	très pour	m1
	plutôt pour	m2
	indécis	m3
	plutôt contre	m4
	très contre	m5

Nombre de mesures obligatoires acceptées

0	z0
1	z1
2	z2
3	z3
4	z4
5	z5

Nombre de situations en public jugées dangereuses

0-1	o1
2	o2
3	o3
4	o4
5-6	o5

Le sida est un péril omniprésent

d'accord	g1
indécis	g2
pas d'accord	g3

Ouvrez le classeur Hahn.stw et observez la façon dont a été constitué le tableau de contingence : la variable "groupe" est croisée avec toutes les autres variables, et on juxtapose ainsi 14 tableaux de contingence portant sur des populations presque identiques (presque, car pour la plupart des questions, il y a quelques non-réponses).

Réalisez une analyse des correspondances sur ce tableau et retrouvez ainsi les résultats de l'auteur :

"L'analyse des correspondances confirme l'existence de deux syndromes nettement distincts, attribuables, avec la prudence qui s'impose, à deux catégories ou milieux, qu'à la suite de Schulze on pourrait appeler "milieu harmoniste" et "milieu autodéterministe".

Notre analyse utilise la dangerosité ressentie du sida comme la variable à décrire, les autres caractéristiques servant d'indices de cette appréciation. Etant donné les trois configurations de la variable à décrire, une solution bidimensionnelle serait théoriquement possible. Mais, puisque le premier axe d'inertie rend compte de 90,25% de la variation, nous négligerons ce deuxième axe.

Graphique et tableau numérique montrent que la vision du sida comme péril a été reportée sur l'ordonnée. On distingue nettement deux groupes, qui approuvent ou rejettent les termes de la question. Ceux qui ne se prononcent pas se situent entre les deux, mais sont enclins le cas échéant à considérer le sida comme une maladie omniprésente et très infectieuse.

À cela correspond la localisation des indicateurs de dispositions (perceptions, réactions) et des repères de morphologie sociale. Les enquêtés considérant le sida comme un péril le jugent très infectieux jusque dans la vie quotidienne (3 situations courantes ou plus jugées contaminantes par un taux supérieur à la moyenne). La maladie est ressentie comme conséquence et punition d'une faute morale; les dispositions d'exclusion se manifestent nettement, et les mesures obligatoires antisida - y compris la généralisation du test obligatoire - rencontrent un taux d'adhésion supérieur à la moyenne. Ceci vérifie nos hypothèses de départ : poussée à l'extrême, la conception du sida comme danger permanent de contamination fait considérer comme porteurs de virus potentiels non seulement les membres des principaux groupes à risque.(donc une minorité), mais tous les étrangers. Les mêmes

enquêtés ressentent la sphère publique comme généralement inquiétante et hostile. Leurs opinions politiques plutôt conservatrices sont attestées par une préférence très nette pour les partis CDU/CSU. Ce groupe comprend une proportion importante de personnes âgées, de niveau social peu élevé, résidant plutôt dans des communes petites ou très petites.

A l'inverse, ceux pour qui le sida n'est pas un péril au sens indiqué ci-dessus, ont pour caractéristique commune de ne pas chercher un risque de contamination là où, en l'état actuel des connaissances, un tel risque n'existe pas. On n'envisage guère la maladie en termes de culpabilité, et on réclame rarement l'exclusion des contaminés ou l'adoption de mesures répressives. Or, ces personnes sont objectivement plus exposées à la contamination.: la fidélité sexuelle est jugée relativement moins importante, le changement de partenaire est relativement fréquent. Les considérations éthico-religieuses passent à l'arrière-plan, la proportion des personnes sans confession est relativement élevée. Politiquement, ce segment se situe majoritairement à gauche du centre, avec une préférence marquée pour les Verts. Morphologiquement, il s'agit d'une population plutôt jeune, étudiante, de niveau social élevé et majoritairement citadine."

### **3.5 Des données d'enquête en "vraie grandeur" : l'enquête "Internet : accès et utilisations au Québec" N° 4 du RISQ**

Consultez le fichier Enquete.pdf et affichez le questionnaire d'enquête dans un navigateur.

Les données se trouvent réparties dans les fichiers enq\_4a.txt à enq\_4i.txt. Il faut commencer par importer chaque fichier dans Statistica (ou dans Excel, qui peut être plus pratique ici), et réassembler les données provenant des 9 fichiers. Il y a au total 8524 observations.

A titre d'exercice, faites-le pour les deux premiers fichiers. Il est ensuite nécessaire de "nettoyer" légèrement le fichier de données obtenu, notamment en rétablissant le statut de "valeur manquante" pour les cases qui contiennent "##".

Pour cela, utiliser l'item de menu Edition - Remplacer. La valeur à rechercher est "##", la valeur de remplacement est laissée vide.

De même, pour certaines questions, c'est la chaîne de caractères "À spécifier :" qui joue le rôle de valeur manquante.

Pour la suite de l'étude, on va travailler sur la feuille de données "Donnees-completes-avecVM.sta" qui contient l'ensemble des données recueillies.

Etudier à l'aide de l'AFC un couple de variables choisi de façon pertinente. Vous serez sans doute conduits à éliminer certaines modalités pour obtenir une représentation interprétable.

## 4 Analyse des Correspondances Multiples

### 4.1 Introduction

L'analyse factorielle des correspondances, vue dans le paragraphe précédent, s'applique à des situations où les individus statistiques sont décrits par *deux* variables nominales. Mais il est fréquent que l'on dispose d'individus décrits par *plusieurs* (deux ou plus) variables nominales ou ordinales. C'est notamment le cas lorsque nos données sont les résultats d'une enquête basée sur des questions fermées. Une extension de l'AFC à ces situations a donc été proposée. Elle est généralement appelée Analyse des Correspondances Multiples ou ACM.

Nous nous plaçons donc dans la situation où nous disposons de  $N$  individus statistiques, décrits par  $q$  variables nominales ou ordinales  $X_1, X_2, \dots, X_q$ . L'ACM vise à mettre en évidence :

- les relations entre les modalités des différentes variables ;
- éventuellement, les relations entre individus statistiques ;
- les relations entre les variables, telles qu'elles apparaissent à partir des relations entre modalités.

### 4.2 Forme des données d'entrée

Selon leur origine, les données sur lesquelles nous nous proposons de faire une ACM peuvent se présenter sous différentes formes.

Imaginons, par exemple, une mini-enquête dans laquelle nous avons posé trois questions à 10 sujets : le sexe (F ou H), le niveau de revenus (M : modeste, E : élevé) et leur préférence sur un sujet donné (3 modalités : A, B ou C). Les données peuvent se présenter sous l'une des formes décrites ci-dessous. Le classeur Mini-ACM.stw contient 5 feuilles de données correspondant à ces 5 formes.

#### 4.2.1 Tableau protocole

	1 Sexe	2 Revenu	3 Preference
s1	F	M	A
s2	F	M	A
s3	F	E	B
s4	F	E	C
s5	F	E	C
s6	H	E	C
s7	H	E	B
s8	H	M	B
s9	H	M	B
s10	H	M	A

### 4.2.2 Tableau d'effectifs

	1	2	3	4
	Sexe	Revenu	reference	Effectif
1	F	M	A	2
2	F	E	B	1
3	F	E	C	2
4	H	E	C	1
5	H	E	B	1
6	H	M	B	2
7	H	M	A	1

### 4.2.3 Tableau disjonctif complet

Le tableau disjonctif complet ou TDC comporte une colonne pour chaque modalité des variables étudiées et une ligne pour chaque individu statistique. Les cellules du tableau contiennent 1 ou 0 selon que l'individu considéré présente la modalité ou non.

	1	2	3	4	5	6	7
	Sexe:F	Sexe:H	Rev:M	Rev:E	Pref:A	Pref:B	Pref:C
s1	1	0	1	0	1	0	0
s2	1	0	1	0	1	0	0
s3	1	0	0	1	0	1	0
s4	1	0	0	1	0	0	1
s5	1	0	0	1	0	0	1
s6	0	1	0	1	0	0	1
s7	0	1	0	1	0	1	0
s8	0	1	1	0	0	1	0
s9	0	1	1	0	0	1	0
s10	0	1	1	0	1	0	0

### 4.2.4 Tableau disjonctif des patrons

En regroupant les lignes identiques du tableau disjonctif complet, on obtient le tableau disjonctif des patrons :

	1	2	3	4	5	6	7
	Sexe:F	Sexe:H	Rev:M	Rev:E	Pref:A	Pref:B	Pref:C
FMA	2	0	2	0	2	0	0
FEB	1	0	0	1	0	1	0
FEC	2	0	0	2	0	0	2
HEC	0	1	0	1	0	0	1
HEB	0	1	0	1	0	1	0
HMB	0	2	2	0	0	2	0
HMA	0	1	1	0	1	0	0

### 4.2.5 Tableau de Burt

L'ACM peut également être réalisée à partir d'une structuration particulière des données, appelée tableau de Burt (TdB). Ce dernier tableau comporte une ligne et une colonne pour chaque modalité des variables étudiées. Chaque cellule du tableau indique le nombre d'individus statistiques qui possèdent en même temps la modalité ligne et la modalité colonne correspondantes. Le tableau de Burt apparaît ainsi comme une juxtaposition de tableaux de contingence des variables prises deux à deux.

	Table Observée (Effectifs) (Protocole dans Classeur2)						
	Table d'Entrée (Lignes x Colonnes) : 7 x 7 (Table de Burt)						
	F	H	M	E	A	B	C
Sexe:F	5	0	2	3	2	1	2
Sexe:H	0	5	3	2	1	3	1
Revenu:M	2	3	5	0	3	2	0
Revenu:E	3	2	0	5	0	2	3
Preference:A	2	1	3	0	3	0	0
Preference:B	1	3	2	2	0	4	0
Preference:C	2	1	0	3	0	0	3

On peut noter qu'il est possible, sans grand problème de passer de l'une des 4 premières structures de données à une autre. De même, le TdB peut être obtenu facilement à partir du tableau disjonctif complet. En revanche, il n'existe pas de moyen simple pour recomposer l'une des 4 premières structures de données à partir du tableau de Burt.

### 4.3 ACM avec Statistica

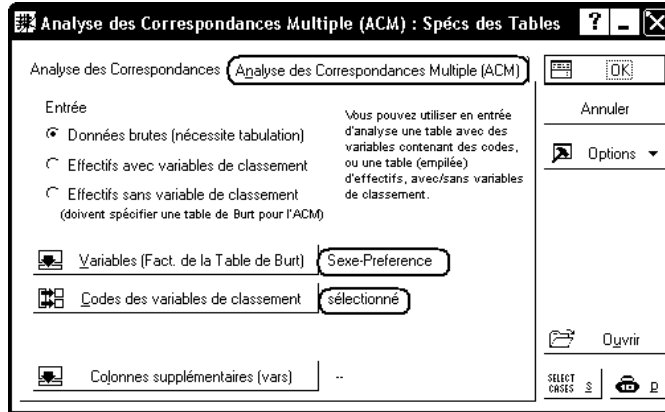
Comme l'indiquent Rouanet et Le Roux :

*Effectuer l'analyse des correspondances multiples, c'est effectuer l'analyse factorielle des correspondances du tableau disjonctif complet, muni des relations  $K < Q >$  (modalités emboîtées dans les questions) et  $I < K < q >$  (individus emboîtés dans les modalités de chaque question).*

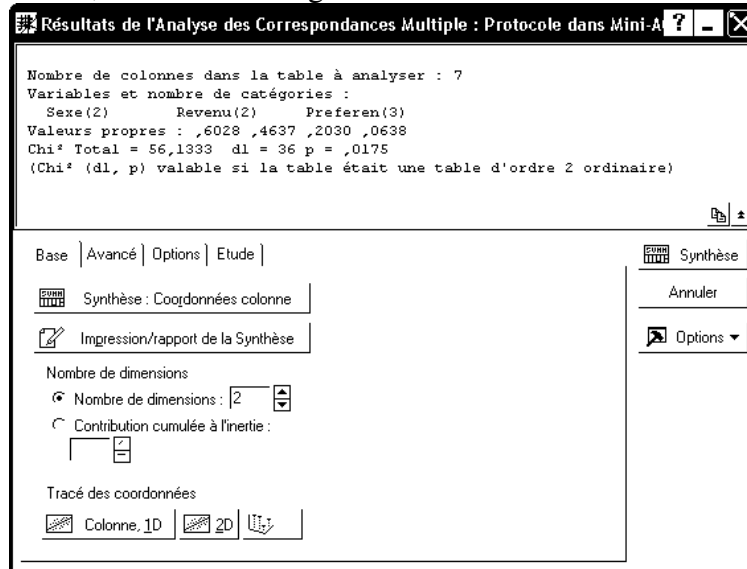
Quelle que soit la forme des données d'entrée, l'ACM sera réalisée à partir du menu Statistiques - Techniques exploratoires multivariées - Analyse des correspondances. Mais, selon la structure des données, c'est l'onglet "Analyse de correspondances" ou l'onglet "Analyse des correspondances multiples (ACM)" qui sera utilisé, selon le tableau suivant :

Format des données	Onglet "Analyse des Correspondances"	Onglet "Analyse des Correspondances Multiple"	Observations
Tableau protocole	Non	Oui	AFC impossible si plus de 2 variables
Tableau d'effectifs	Non	Oui	AFC impossible si plus de 2 variables
Tableau Disjonctif Complet	Oui	Non	
Tableau Disjonctif des patrons	Oui	Non	
Tableau de Burt	Oui	Oui	Les deux analyses ne fournissent pas les mêmes résultats

Réalisons, par exemple, une ACM sur le tableau protocole. Après avoir déclaré cette feuille de données comme 'feuille active', on sélectionne l'onglet "Analyse des correspondances multiple" et on complète le premier dialogue comme suit :



Une fois ce dialogue validé, un second dialogue s'affiche :

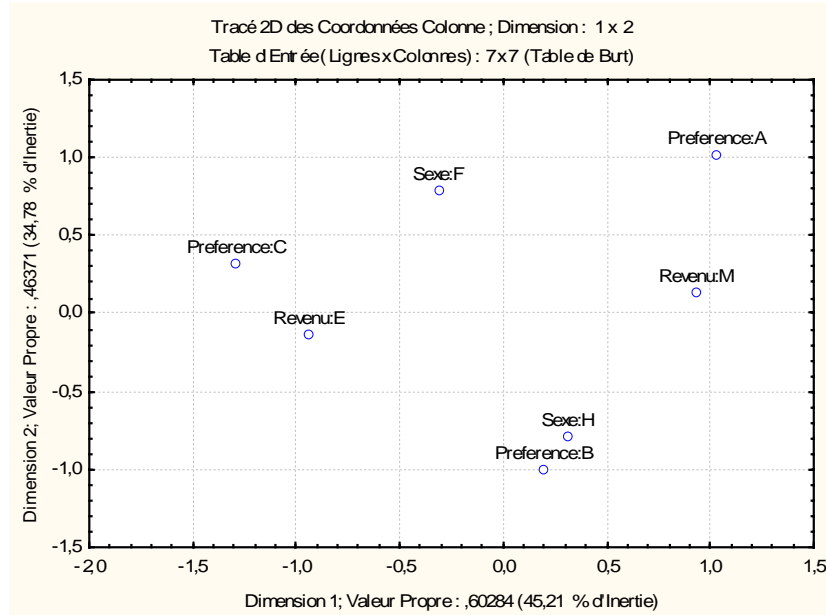


Le bouton "Effectifs Observés de l'onglet "Etude" permet d'obtenir un tableau similaire au tableau de Burt. Les pourcentages ligne, pourcentages colonne, khi-2, etc utilisent ce dernier tableau.

L'onglet "Avancé" permet d'obtenir les autres résultats :

Valeurs Propres et Inertie de toutes les Dimensions (Protocole dans Mini-ACM-v7.stw) Table d'Entrée (Lignes x Colonnes) : 7 x 7 (Table de Burt) Inertie Totale = 1,3333					
Nombre de Dims.	ValSing.	ValProp.	%age Inertie	%age Cumulé	Chi²
1	0,7764	0,6028	45,2128	45,2128	25,3794
2	0,6810	0,4637	34,7781	79,9909	19,5221
3	0,4505	0,2030	15,2219	95,2128	8,5446
4	0,2526	0,0638	4,7872	100,0000	2,6872

Coordonnées Colonne et Contributions à l'Inertie (Protocole dans Mini-ACM-v7.stw) Table d'Entrée (Lignes x Colonnes) : 7 x 7 (Table de Burt) Inertie Totale = 1,3333										
NomLigne	Ligne Numéro	Coord. Dim.1	Coord. Dim.2	Masse	Qualité	Inertie Relative	Inertie Dim.1	cosinus Dim.1	Inertie Dim.2	cosinus Dim.2
Sexe:F	1	-0,311	0,788	0,167	0,718	0,125	0,027	0,097	0,223	0,621
Sexe:H	2	0,311	-0,788	0,167	0,718	0,125	0,027	0,097	0,223	0,621
Revenu:M	3	0,938	0,138	0,167	0,899	0,125	0,243	0,880	0,007	0,019
Revenu:E	4	-0,938	-0,138	0,167	0,899	0,125	0,243	0,880	0,007	0,019
Preference:A	5	1,032	1,024	0,100	0,906	0,175	0,177	0,456	0,226	0,450
Preference:B	6	0,193	-1,007	0,133	0,701	0,150	0,008	0,025	0,292	0,677
Preference:C	7	-1,288	0,319	0,100	0,755	0,175	0,275	0,711	0,022	0,044



Bien que l'exemple ne comporte qu'un petit nombre d'observations, on remarque la proximité des modalités Préférence:B et Sexe:H, de même que l'opposition Préférence C, revenu E d'une part, Préférence A, Revenu M d'autre part selon le premier axe.

On note également que l'origine du repère est le milieu du segment joignant les deux modalités de la variable "Sexe", et aussi le milieu du segment joignant les deux modalités de la variable "Revenu". En effet, ces deux variables ont seulement deux modalités (d'où l'alignement de l'origine avec les modalités) et les deux modalités sont équiprobables (d'où la propriété du milieu).

On pourra recommencer l'étude en utilisant les autres feuilles de données, et on obtiendra ainsi des résultats analogues. Seule l'étude à partir du tableau disjonctif complet permet, éventuellement, de placer les individus sur le graphique.

L'étude menée à partir du tableau de Burt mérite un commentaire particulier. En effet, dans un exposé théorique sur l'ACM, tels que ceux de [Crucianu] ou de [Rouanet, Le Roux], l'analyse du tableau de Burt est distinguée de celle du TDC ou du tableau disjonctif des patrons. Il est notamment indiqué que les valeurs propres produites par cette analyse sont les carrés des valeurs propres précédentes, et que le  $\Phi^2$  du tableau de Burt n'est pas celui du TDC. Cependant, les représentations graphiques produites (limitées aux seules modalités) peuvent être interprétées de façon analogue.

Or, avec Statistica, on constate que l'on obtient, pour les modalités, des résultats identiques aux précédents. En particulier, les valeurs propres sont celles qui ont indiquées plus haut.

En revanche, nous pouvons effectuer une AFC à l'aide de l'onglet "Analyse des correspondances", en spécifiant le tableau de Burt comme tableau de contingence. On retrouve alors les résultats indiqués dans les exposés théoriques. Par exemple, le tableau des valeurs propres est alors donné par :

Nombre de Dims.	Valeurs Propres et Inertie de toutes les Dimensions (Tableau de Burt dans Mini-ACM-v7.stw)				
	ValSing.	ValProp.	%age Inertie	%age Cumulé	Chi <sup>2</sup>
	Table d'Entrée (Lignes x Colonnes) : 7 x 7 Inertie Totale = ,62370 Chi <sup>2</sup> = 56,133 dl = 36 p = ,01747				
1	0,6028	0,3634	58,2668	58,2668	32,7071
2	0,4637	0,2150	34,4755	92,7423	19,3523
3	0,2030	0,0412	6,6045	99,3468	3,7073
4	0,0638	0,0041	0,6532	100,0000	0,3667
5	0,0000	0,0000	0,0000	100,0000	0,0000
6	0,0000	0,0000	0,0000	100,0000	0,0000

#### 4.4 Quelques règles d'interprétation

On cherchera d'une part à interpréter les oppositions entre modalités (ou entre groupes d'individus, si l'étude porte sur le TDC), et d'autre part à interpréter les proximités entre modalités.

L'interprétation des proximités entre les modalités devra tenir compte de la remarque suivante :

- Si deux modalités *d'une même variable* sont proches, cela signifie que les individus qui possèdent l'une des modalités et ceux qui possèdent l'autre sont globalement similaires *du point de vue des autres variables* ;
- Si deux modalités *de deux variables différentes* sont proches, cela peut signifier que ce sont globalement les mêmes individus qui possèdent l'une et l'autre.

Nous pouvons, comme en AFC, nous intéresser aux profils ligne et colonne, aux taux de liaison et au  $\Phi^2$  du tableau disjonctif complet, vu comme un tableau de contingence. Le nombre de lignes de ce tableau est égal au nombre d'individus statistiques étudiés. Cependant, nous avons vu que la métrique du  $\Phi^2$ , utilisée pour l'AFC, possède la propriété d'équivalence distributionnelle : si on regroupe deux lignes correspondant au même patron de réponses, on ne change rien aux autres profils lignes, ni aux autres profils colonnes. Autrement dit, on retrouvera les mêmes résultats en effectuant une AFC sur le tableau disjonctif des patrons.

Comme en AFC, on peut calculer des fréquences, des fréquences lignes, des fréquences colonnes et des profils lignes et profils colonnes moyens.

L'élément le plus facile à interpréter est le profil colonne moyen : ce sont les fréquences des différents patrons de réponses dans la population étudiée.

##### 4.4.1 Distances entre profils lignes

En AFC, nous avons donné les formules permettant de calculer les distances entre deux profils lignes ou entre deux profils colonnes. La distance utilisée est la *métrique du  $\Phi^2$* . Ici, compte tenu de la structure particulière du tableau de contingence utilisé, les formules indiquées deviennent :

$$d_{\Phi^2}^2(L_i, L_{i'}) = \frac{1}{Q} \sum_k \frac{(\delta_{ik} - \delta_{i'k})^2}{f_{\cdot k}}$$

Notations utilisées :  $L_i$  et  $L_{i'}$  désignent deux patrons,  $Q$  est le nombre de questions.  $\delta_{ik}$  prend la valeur 1 si la modalité  $k$  fait partie du patron  $i$ , et la valeur 0 sinon. Enfin,  $f_{\cdot k}$  est la fréquence de la modalité  $k$  dans la population.

Autrement dit, deux individus (ou deux patrons) sont d'autant plus éloignés que leurs réponses diffèrent pour un plus grand nombre de questions et pour des modalités rares.

La distance d'un patron au profil ligne moyen est :

$$d_{\Phi^2}^2(O, L_i) = \left( \frac{1}{Q} \sum_k \frac{\delta_{ik}}{f_{\bullet k}} \right) - 1$$

Autrement dit, un patron sera d'autant plus loin de l'origine qu'il fait intervenir des modalités plus rares.

La contribution (absolue) d'un patron à la variance du nuage est obtenue en multipliant la distance précédente par la fréquence du patron dans la population.

#### 4.4.2 Distances entre profils colonnes

La distance entre les modalités k et k' est donnée par :

$$d_{\Phi^2}^2(M_k, M_{k'}) = \frac{1}{f_{\bullet k}} + \frac{1}{f_{\bullet k'}} - 2 \frac{f_{kk'}}{f_{\bullet k} f_{\bullet k'}}$$

où  $f_{kk'}$  est la fréquence de la combinaison de modalités k et k'.

La distance d'une modalité au profil colonne moyen est donnée par :

$$d_{\Phi^2}^2(O, M_k) = \frac{1}{f_{\bullet k}} - 1$$

Autrement dit, une modalité sera d'autant plus loin du profil moyen que sa fréquence est faible.

La contribution absolue d'une modalité à la variance du nuage de points est :

$$Cta(M_k) = \frac{1 - f_{\bullet k}}{Q}$$

#### 4.4.3 Taux de liaison et Phi-2

Pour le tableau disjonctif complet, ou le tableau disjonctif des patrons, considérés comme des tableaux de contingence, le coefficient Phi-2 vaut :

$$\Phi^2 = \frac{K}{Q} - 1$$

où K désigne le nombre de modalités et Q le nombre de questions

Dans notre exemple, on a : K=7, Q=3, et donc :  $\Phi^2 = \frac{7}{3} - 1 = 1,33$ .

### 4.5 Autres exemples d'ACM

Les autres exemples d'ACM que nous traiterons sont données à l'aide d'un tableau de Burt. En effet, c'est généralement sous cette forme que l'on trouve des données susceptibles de servir de base à un exercice.

#### 4.5.1 Le cas "Aspirations des Français"

Ouvrez le classeur Aspi.stw. La présentation du cas, rappelée dans un rapport contenu dans le classeur est la suivante :

Source : Morineau A., Morin S., Pratique du traitement des enquêtes - Exemple d'utilisation du système SPAD, Cisia-Ceresta, Montreuil, 2000

On travaille sur des données extraites d'une enquête d'opinion réalisée en 1978, concernant les conditions de vie et les aspirations des Français.

Les questions prises en compte ici, et leurs modalités, sont les suivantes :

1- Sexe de la personne interrogée :

masc : masculin

femi : féminin

2- Possédez-vous des valeurs mobilières

vmo1 : oui

vmo2 : non

3- Taille d'agglomération

agg1 : moins de 2000 h

agg2 : de 2000 à 20000 h

agg3 : de 20000 à 100000 h

agg4 : plus de 100000h

agg5 : Paris

4- Diplôme de l'enquêté :

die1 : aucun

die2 : CEP ou fin d'études

die3 : BEPC - BE - BEPS

die4 : bac - brevet sup.

die5 : université, gde école

5- Statut du logement

slo1 : en accession

slo2 : propriétaire

slo3 : locataire

slo4 : logé gratuit, autre

6- Age de l'enquêté

agc1 : moins de 25 ans

agc2 : 25 à 34 ans

agc3 : 35 à 49 ans

agc4 : 50 à 64 ans

agc5 : plus de 65 ans

7- Type d'emploi

emp1 : ouvriers

emp2 : employés


emp3 : cadres

emp4 : autres

empNR : non réponse

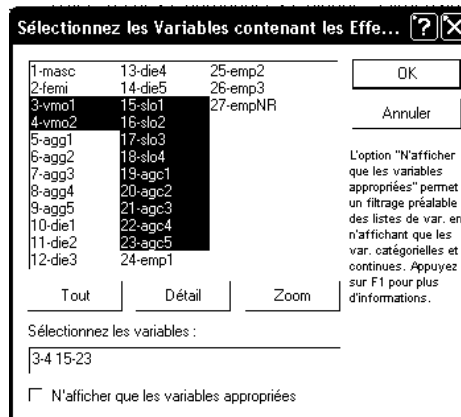
Remarque : pour une ACM sur la totalité des 27 modalités du TDB, les auteurs retiennent 5 axes principaux.

Faites tout d'abord une ACM sur la totalité du tableau de Burt (27 modalités - remarquez que seules 4 modalités de la variable "Type d'emploi" sont présentes.

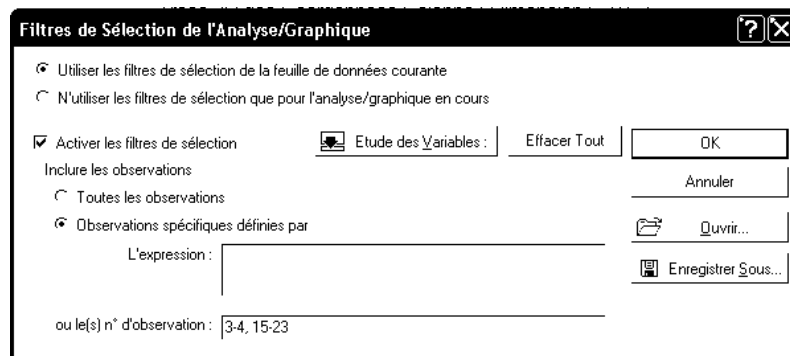
**Remarque** : le graphique ainsi obtenu est assez peu lisible. Il est cependant possible de l'améliorer en utilisant l'outil "Balayage/Habillage" :  . A l'aide de cet outil, il est par exemple possible de supprimer certains points qui se superposent au centre du graphique. Attention cependant à ce que le graphique conserve une certaine honnêteté intellectuelle !

Réalisez ensuite une ACM en ne prenant en compte que certaines variables, par exemple, la variable 2 (valeurs mobilières), la variable 5 (statut du logement) et la variable 6 (âge de l'enquêté). Pour cela :

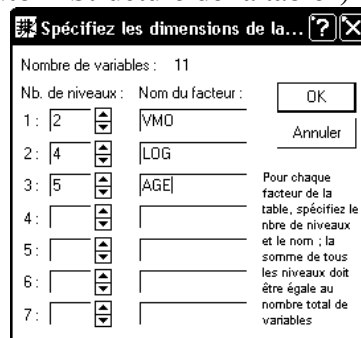
- Sélectionnez les variables comme suit :



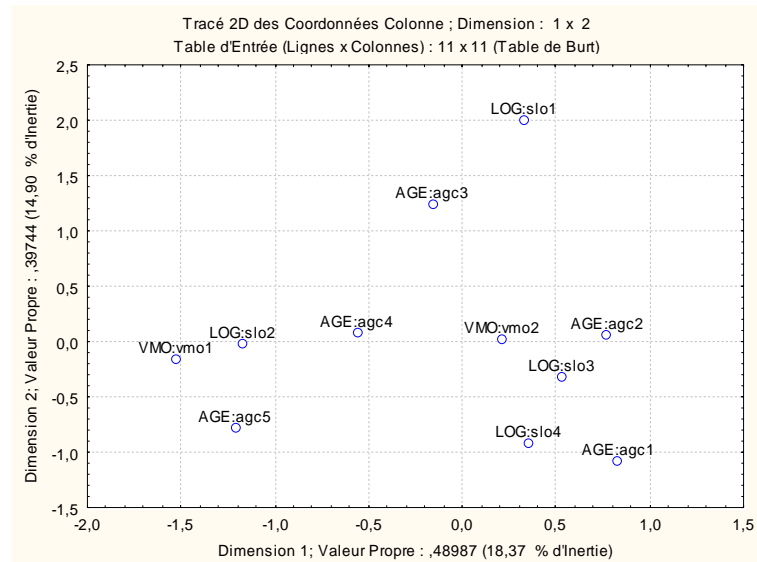
- Sélectionnez ensuite les observations correspondantes, par exemple en les désignant par leurs numéros. Pour cela, cliquez sur le bouton "Select Cases" et complétez le dialogue comme suit :



- Structurez enfin les variables (bouton "Structure de la table") de la façon suivante :



On obtient ainsi le graphique suivant :



La possession de valeurs mobilières est ainsi plutôt associée à l'occupation d'un logement en propriété, et à une personne relativement âgée (agc4, agc5), alors que la non-possession est plutôt le fait de personnes jeunes, locataires. L'âge agc3 est dans une certaine mesure associé à l'accession à la propriété alors que le dernier statut du logement est plutôt le fait des moins de 25 ans (qui, par ailleurs, ne possèdent généralement pas de valeurs mobilières).

#### 4.5.2 Le cas "Avignon"

Source : Croutsche, J.-J., Pratiques statistiques en gestion et études de marchés, Editions ESKA, Paris, 1997

Une enquête sur la fréquentation du centre ville d'Avignon. On trouvera ci-dessous le texte d'une partie des questions posées, ainsi que le codage des modalités de réponse.

- 1- Combien de fois par mois allez-vous dans le centre ville pour faire des achats ?
  - a1 : Plus de 3 fois par mois
  - a2 : de 2 à 3 fois
  - a3 : de 1 à 2 fois
  - a4 : Autre
- 2- Votre fréquentation du centre ville est-elle plus ou moins importante qu'il y a 5 ans ?
  - f1 : Beaucoup moins importante
  - f2 : Un peu moins importante
  - f3 : Identique
  - f4 : Un peu plus importante
  - f5 : Beaucoup plus importante
- 3-
- 4-
- 5- Etes-vous satisfait de la propreté du centre ville ?
  - p1 : très satisfait
  - p2 : satisfait
  - p3 : moyennement satisfait
  - p4 : peu satisfait
  - p5 : très peu satisfait
- 6- Que pensez-vous de la sécurité dans le centre ville ?
  - s1 : Très faible
  - s2 : Faible
  - s3 : Normale

- s4 : Importante
- s5 : Très importante

7- Si vous observez des problèmes de sécurité : vous arrive-t-il de ne pas vous rendre dans le centre ville à cause de ce problème ?

- r1 : oui
- r2 : non

8-

9-

10-

11- Où habitez-vous ?

- h1 : Avignon intra-muros
- h2 : Avignon extra-muros
- h3 : autre

12-

13- Dans quelle tranche d'âge vous situez-vous ?

- â1 : 15-19 ans
- â2 : 20-30 ans
- â3 : 31-40 ans
- â4 : 41-50 ans
- â5 : 51-60 ans
- â6 : Plus de 60 ans

14-

Dans le classeur Avignon.stw se trouvent diverses feuilles de données contenant les tableaux de Burt obtenus en sélectionnant 3 ou 4 des items du questionnaire. Analysez chacun des aspects ainsi définis à l'aide d'une ACM.

## 5 Classification Ascendante Hiérarchique

### 5.1 Introduction

Classifier, c'est regrouper entre eux des objets similaires selon tel ou tel critère. Les diverses techniques de classification (ou d'"analyse typologique", de "taxonomie", ou "taxinomie" ou encore "analyse en clusters" (amas)) visent toutes à répartir  $n$  individus, caractérisés par  $p$  variables  $X_1, X_2, \dots, X_p$  en un certain nombre  $m$  de sous-groupes aussi homogènes que possible.

On distingue deux grandes familles de techniques de classification :

- La classification non hiérarchique ou partitionnement, aboutissant à la décomposition de l'ensemble de tous les individus en  $m$  ensembles disjoints ou classes d'équivalence ; le nombre  $m$  de classes est fixé.
- La classification hiérarchique : pour un niveau de précision donné, deux individus peuvent être confondus dans un même groupe, alors qu'à un niveau de précision plus élevé, ils seront distingués et appartiendront à deux sous-groupes différents.

Remarques. Ces méthodes jouent un rôle un peu à part dans l'univers des méthodes statistiques. En effet :

- L'aspect inférentiel est ici inexistant ;
- Il existe un grand nombre de variantes de ces méthodes, et on peut être amené à appliquer plusieurs de ces méthodes sur un même jeu de données, jusqu'à obtenir une classification "qui fasse sens" ;

- Au contraire des méthodes factorielles, l'accent est souvent mis sur les  $n$  individus et non sur les  $p$  variables qui les décrivent.

## 5.2 Les 4 étapes de la méthode

### 5.2.1 Choix des variables représentant les individus

Dans le cas où les données observées sont les valeurs de  $p$  variables numériques sur  $n$  individus, on pourra choisir d'effectuer une classification des individus, ou une classification des variables. On peut choisir, par exemple, de retenir certains "traits" des individus (autrement dit certaines variables qui ont servi à les décrire) et réaliser la classification sur les individus décrits par ce choix de variables.

On peut noter qu'il revient au même par exemple :

- de réaliser la CAH des individus à partir de  $p$  variables centrées réduites ;
- de réaliser la CAH des individus à partir des  $p$  facteurs obtenus à l'aide d'une ACP normée sur les variables précédentes.

Toutefois, il peut être intéressant de réaliser la CAH à partir des  $q$  premiers facteurs ( $q < p$ ). Cela a pour effet d'éliminer une partie des variations entre individus, qui correspond en général à des fluctuations aléatoires, c'est-à-dire à un "bruit statistique".

Dans le cas où les données observées sont représentées par un tableau de contingence, c'est-à-dire sont les valeurs de 2 variables nominales sur  $n$  individus, on pourra effectuer une CAH des modalités-lignes par exemple, à partir des coordonnées lignes obtenues par une AFC. On pourra, de même, réaliser une CAH des modalités-colonnes.

Enfin, si les données observées sont les valeurs de  $p$  variables nominales sur  $n$  individus, on pourra effectuer une CAH des individus en partant du tableau disjonctif complet, ou en utilisant les coordonnées des individus obtenues par une ACM. On pourra également traiter les modalités comme dans le cas d'une AFC.

### 5.2.2 Choix d'un indice de dissimilarité

De nombreuses mesures de la "distance" entre individus ont été proposées. Le choix d'une (ou plusieurs) d'entre elles dépend des données étudiées. Statistica nous propose les mesures suivantes :

- Distance Euclidienne. C'est probablement le type de distance le plus couramment utilisé. Il s'agit simplement d'une distance géométrique dans un espace multidimensionnel.

$$d(I_i, I_j) = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$$

- Distance Euclidienne au carré. On peut élever la distance euclidienne standard au carré afin de "sur-pondérer" les objets atypiques (éloignés).

$$d(I_i, I_j) = \sum_k (x_{ik} - x_{jk})^2$$

- Distance du City-block (Manhattan) :

$$d(I_i, I_j) = \sum_k |x_{ik} - x_{jk}|$$

- Distance de Tchebychev :

$$d(I_i, I_j) = \text{Max} |x_{ik} - x_{jk}|$$

- Distance à la puissance.

$$d(I_i, I_j) = \left( \sum_k |x_{ik} - x_{jk}|^p \right)^{1/r}$$

- Percent disagreement. Cette mesure est particulièrement utile si les données des dimensions utilisées dans l'analyse sont de nature catégorielle.

$$d(I_i, I_j) = \frac{\text{Nombre de } x_{ik} \neq x_{jk}}{K}$$

- 1- r de Pearson : calculée à partir du coefficient de corrélation, à l'aide de la formule :

$$d(I_i, I_j) = 1 - r_{ij}$$

### 5.2.2.1 Indices de dissimilarité et distances

On peut également utiliser d'autres indices de dissimilarité puisque Statistica permet d'effectuer la classification à partir du tableau des scores de dissimilarités entre individus. En fait, un indice de dissimilarité doit simplement satisfaire les conditions suivantes :

- non-négativité :  $d(I_i, I_j) \geq 0$
- symétrie :  $d(I_i, I_j) = d(I_j, I_i)$
- normalisation :  $d(I_i, I_i) = 0$

Un indice de dissimilarité est une "vraie" distance, s'il vérifie également l'inégalité triangulaire :

$$d(I_i, I_j) \leq d(I_i, I_k) + d(I_k, I_j).$$

La plupart des "distances" proposées par Statistica sont de véritables distances.

De nombreux indices de dissimilarité (ou au contraire de similarité) ont été proposés dans le cas de variables qualitatives (à deux modalités, ou après codage disjonctif). Par exemple, si les individus sont décrits par K variables dichotomiques (oui/non), on peut introduire :

$$a_{ij} = \text{Nombre co-occurrences entre les individus } i \text{ et } j$$

$$d_{ij} = \text{Nombre co-absences entre les individus } i \text{ et } j$$

$$b_{ij} = \text{Nombre d'attributs présents chez } i \text{ et absents chez } j$$

$$c_{ij} = \text{Nombre d'attributs absents chez } i \text{ et présents chez } j$$

On peut proposer par exemple, comme indice de dissimilarité :

$$d(I_i, I_j) = \sqrt{b_{ij} + c_{ij}}$$

ou au contraire, comme indice de similarité :

$$s(I_i, I_j) = \frac{a_{ij}}{a_{ij} + b_{ij} + c_{ij}}$$

Un indice de similarité peut être converti en distance par la relation :

$$d(I_i, I_j) = s_{\max} - s(I_i, I_j)$$

### 5.2.3 Choix d'un indice d'agrégation

L'application de la méthode suppose également que nous fassions le choix d'une "distance" entre classes. Là encore, de nombreuses solutions existent. Il faut noter que ces solutions permettent toutes de calculer la distance entre deux classes quelconques sans avoir à recalculer celles qui existent entre les individus composant chaque classe.

Les choix proposés par Statistica sont les suivants :

- Saut minimum ou "single linkage" (distance minimum). C'est celle que nous avons utilisée ci-dessus.

- Diamètre ou "complete linkage" (distance maximum). Dans cette méthode, les distances entre classes sont déterminées par la plus grande distance existant entre deux objets de classes différentes (c'est-à-dire les "voisins les plus éloignés").

$$D(A,B) = \max_{I \in A} \max_{J \in B} d(I,J)$$

- Moyenne non pondérée des groupes associés. Ici, la distance entre deux classes est calculée comme la moyenne des distances entre tous les objets pris dans l'une et l'autre des deux classes différentes.

$$D(A,B) = \frac{1}{n_A n_B} \sum_{I \in A, J \in B} d(I,J)$$

- Moyenne pondérée des groupes associés. La moyenne précédente est étendue à l'ensemble des paires d'objets trouvées dans la réunion des deux classes.

$$D(A,B) = \frac{1}{(n_A + n_B)(n_A + n_B - 1)} \sum_{I, J \in A \cup B} d(I,J)$$

- Centroïde non pondéré des groupes associés. Le centroïde d'une classe est le point moyen d'un espace multidimensionnel, défini par les dimensions. Dans cette méthode, la distance entre deux classes est déterminée par la distance entre les centroïdes respectifs.

- Centroïde pondéré des groupes associés (médiane). Cette méthode est identique à la précédente, à la différence près qu'une pondération est introduite dans les calculs afin de prendre en compte les tailles des classes (c'est-à-dire le nombre d'objets contenu dans chacune).

- Méthode de Ward (méthode du moment d'ordre 2). Cette méthode se distingue de toutes les autres en ce sens qu'elle utilise une analyse de la variance approchée afin d'évaluer les distances entre classes. En résumé, cette méthode tente de minimiser la Somme des Carrés (SC) de tous les couples (hypothétiques) de classes pouvant être formés à chaque étape. Les indices d'agrégation sont recalculés à chaque étape à l'aide de la règle suivante : si une classe M est obtenue en regroupant les classes K et L, sa distance à la classe J est donnée par :

$$D(M,J) = \frac{(N_J + N_K)D(K,J) + (N_J + N_L)D(L,J) - N_J D(K,L)}{N_J + N_K + N_L}$$

La méthode de Ward se justifie bien lorsque la "distance" entre les individus est le carré de la distance euclidienne. Choisir de regrouper les deux individus les plus proches revient alors à choisir la paire de points dont l'agrégation entraîne la diminution minimale de l'inertie du nuage. Le calcul des nouveaux indices entre la paire regroupée et les points restants revient alors à remplacer les deux points formant la paire par leur point moyen, affecté du poids 2.

## 5.2.4 Algorithme de classification et résultat produit

### 5.2.4.1 L'algorithme de classification

La classification proprement dite peut être décrite de la manière suivante :

Étape 1 : il y a n éléments à classer (qui sont les n individus);

Étape 2 : on construit la matrice de distances entre les n éléments et l'on cherche les deux plus proches, que l'on agrège en un nouvel élément. On obtient une première partition à n-1 classes;

Étape 3 : on construit une nouvelle matrice des distances qui résultent de l'agrégation, en calculant les distances entre le nouvel élément et les éléments restants (les autres distances sont inchangées).

On se trouve dans les mêmes conditions qu'à l'étape 1, mais avec seulement  $(n-1)$  éléments à classer et en ayant choisi un critère d'agrégation. On cherche de nouveau les deux éléments les plus proches, que l'on agrège. On obtient une deuxième partition avec  $n-2$  classes et qui englobe la première;

Étape  $m$  : on calcule les nouvelles distances, et l'on réitère le processus jusqu'à n'avoir plus qu'un seul élément regroupant tous les objets et qui constitue la dernière partition.

### 5.2.4.2 Hiérarchie de classes et partition de l'ensemble des individus

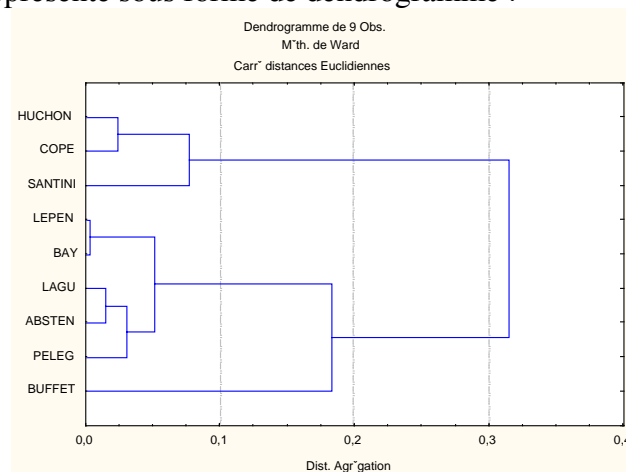
Opérer une classification, c'est définir une partition de l'ensemble des individus, c'est -à-dire, définir un ensemble de parties, ou classes de l'ensemble  $I$  des individus telles que :

- toute classe soit non vide
- deux classes distinctes sont disjointes
- tout individu appartient à une classe.

Le résultat d'une CAH n'est pas une partition de l'ensemble des individus. C'est une hiérarchie de classes telles que :

- toute classe est non vide
- tout individu appartient à une (et même plusieurs) classes
- deux classes distinctes sont disjointes, ou vérifient une relation d'inclusion (l'une d'elles est incluse dans l'autre)
- toute classe est la réunion des classes qui sont incluses dans elle.

Ce résultat est souvent représenté sous forme de dendrogramme :



Sur la figure ci-dessus, l'axe vertical indique les individus statistiques qui ont été rassemblés pour former les classes, tandis que la graduation de l'axe horizontal indique la distance séparant les deux classes qui ont été rassemblées une étape donnée.

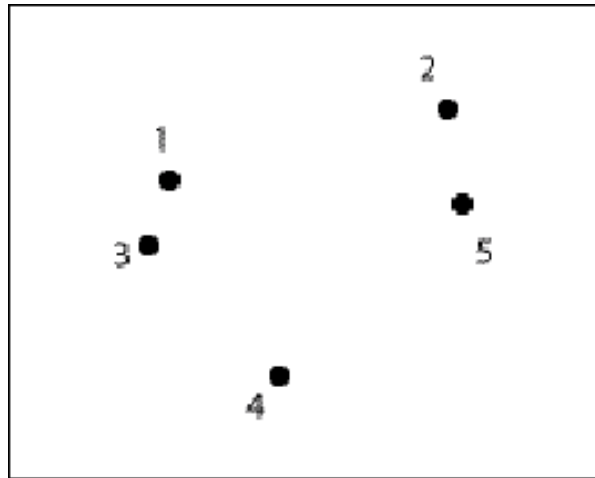
### 5.2.4.3 Choix d'une partition à partir de la hiérarchie des classes

Le dendrogramme nous indique l'ordre dans lequel les agrégations successives ont été opérées. Il nous indique également la valeur de l'indice d'agrégation à chaque niveau d'agrégation. Il est généralement pertinent d'effectuer la coupure après les agrégations correspondant à des valeurs peu élevées de l'indice et avant les agrégations correspondant à des valeurs élevées. En coupant l'arbre au niveau d'un saut important de cet indice, on peut espérer obtenir une partition de bonne qualité car les individus regroupés en-dessous de la coupure étaient proches, et ceux regroupés après la coupure sont éloignés.

### 5.3 CAH "à la main"

Le dessin suivant représente 5 objets "en vraie grandeur". La distance utilisée entre les objets est la distance euclidienne (mesurée au double-décimètre). L'indice d'agrégation est celui du "saut minimal".

Réalisez une CAH sur ces données :



### 5.4 La CAH avec Statistica

#### 5.4.1 Un exemple de CAH effectué à partir d'un tableau de contingence

Source : Lebart L., Morineau A., Piron M. Statistique Exploratoire Multidimensionnelle.

L'exemple concerne l'analyse d'un tableau de contingence qui croise 8 professions et catégories socioprofessionnelles (PCS) et 6 types de médias pour un échantillon de 12 388 "contacts média" relatifs à 4433 personnes interrogées. L'individu statistique sera pour nous le "contact média" et non la personne interrogée dans l'enquête. Les données sont extraites de l'Enquête Budget-temps Multimédia 1991-1992 du CESP.

Afin d'interpréter plus efficacement les représentations obtenues, on projettera en éléments supplémentaires certaines autres caractéristiques de la population enquêtée telles que le sexe, l'âge, le niveau d'instruction.

Tables de contingence croisant les types de contacts-média (colonnes) avec professions, sexe, âge, niveau d'éducation (lignes).

	Radio	Tél.	Quot.N.	Quot R.	P.Mag.	P.TV
Professions						
Agriculteur	96	118	2	71	50	17
Petit patron	122	136	11	76	49	41
Prof. Cad. S.	193	184	74	63	103	79
Prof. interm	360	365	63	145	141	184
Employé	511	593	57	217	172	306
Ouvrier qual	385	457	42	174	104	220
Ouvrier n-q	156	185	8	69	42	85
Inactif	1474	1931	181	852	642	782

Nous disposons des tables de contingence suivantes (cf. tableau) : On trouve, à l'intersection de la ligne i et de la colonne j le nombre  $k_{ij}$  d'individus appartenant à la catégorie i et ayant eu la veille

(un jour de semaine) au moins un contact avec le type de média j. Une personne interrogée pouvant avoir des contacts avec plusieurs médias, les valeurs en ligne représentent des "nombres de contacts".

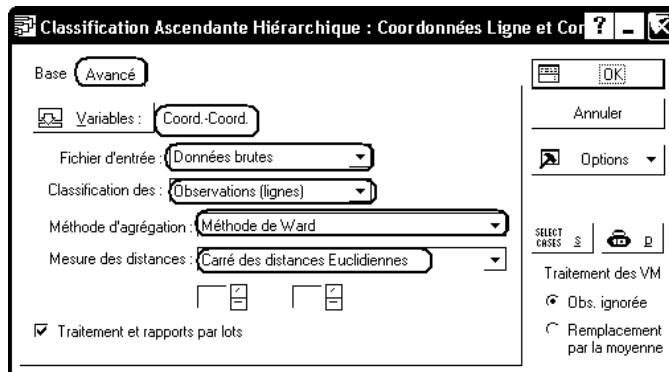
Nous nous proposons de réaliser une CAH sur les professions à partir de ce tableau. Comme nous l'avons vu dans le paragraphe sur l'AFC, la "distance" pertinente entre deux lignes du tableau est la distance du khi-2, ou, ce qui revient au même, le carré de la distance euclidienne entre les images des modalités lignes obtenues par AFC.

Dans un premier temps, ouvrez le classeur Contacts-Medias.stw et réalisez une AFC en calculant les coordonnées lignes et colonnes sur tous les facteurs.

Rendez ensuite active la feuille de données contenant les résultats relatifs aux lignes.

Utilisez ensuite le menu Statistiques - Techniques Exploratoires Multivariées - Classifications .

On choisit ici comme mesure des distances, le carré des distances euclidiennes. Cela revient à mesurer la distance entre deux lignes à l'aide de la distance du khi-2 (propriété de l'AFC). L'indice d'agrégation choisi est celui calculé par la méthode de Ward.



On obtient ainsi les résultats suivants :

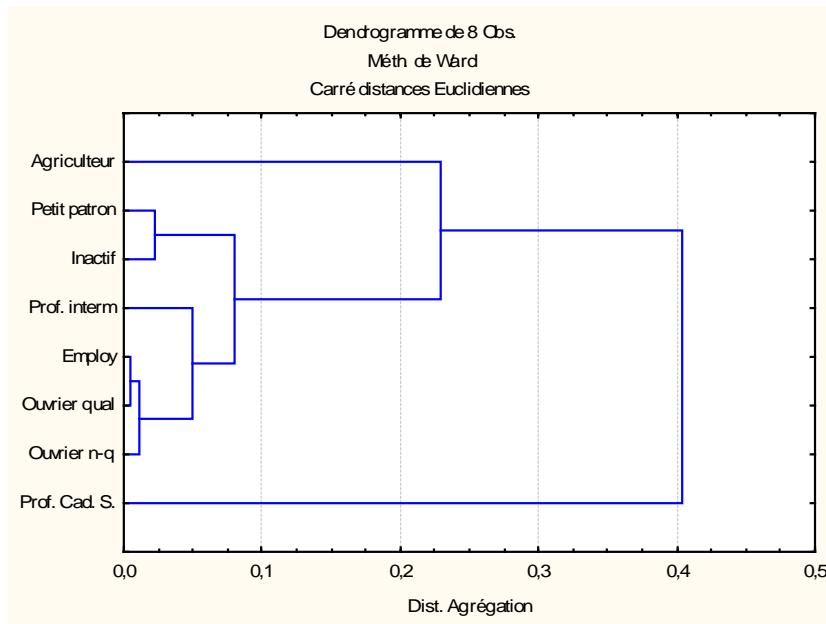
- un tableau donnant les étapes de la classification :

Agrégation Finale (Coordonnées Ligne et Contributions à l'Inertie (Contacts-medias.sta dans Classeur4) dans Classeur Méth. de Ward Carré distances Euclidiennes								
distance agrégat.	Objet # 1	Objet # 2	Objet # 3	Objet # 4	Objet # 5	Objet # 6	Objet # 7	Objet # 8
,0041508	Employé	Ouvrier qual						
,0120153	Employé	Ouvrier qual	Ouvrier n-q					
,0225031	Petit patron	Inactif						
,0496726	Prof. interm	Employé	Ouvrier qual	Ouvrier n-q				
,0805143	Petit patron	Inactif	Prof. interm	Employé	Ouvrier qual	Ouvrier n-q		
,2296203	Agriculteur	Petit patron	Inactif	Prof. interm	Employé	Ouvrier qual	Ouvrier n-q	
,4046085	Agriculteur	Petit patron	Inactif	Prof. interm	Employé	Ouvrier qual	Ouvrier n-q	Prof. Cad. S.

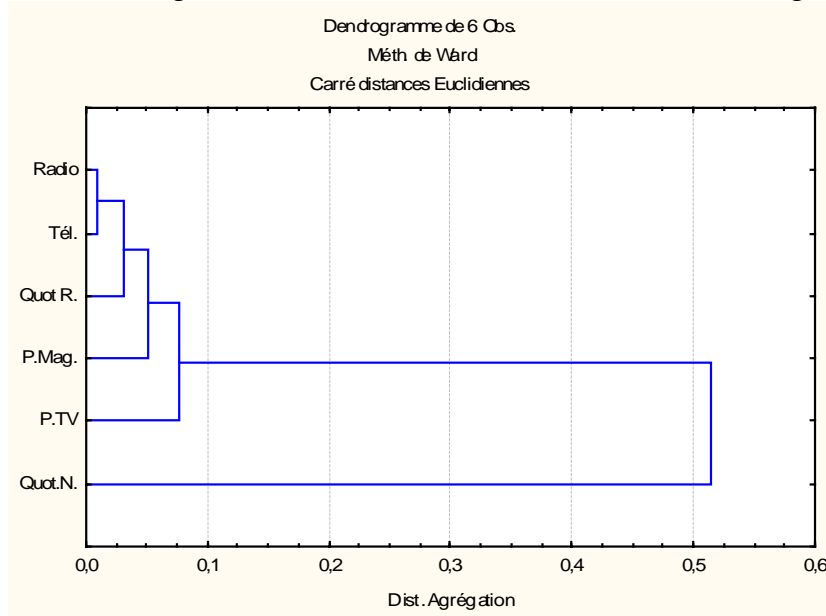
- Le tableau des distances entre individus :

N° Obs.	Agriculteur	Petit patron	Prof. Cad. S.	Prof. interm	Employé	Ouvrier qual	Ouvrier n-q	Inactif
Agriculteur	0,00	0,04	0,42	0,19	0,19	0,19	0,17	0,10
Petit patron	0,04	0,00	0,26	0,06	0,07	0,06	0,06	0,02
Prof. Cad. S.	0,42	0,26	0,00	0,12	0,22	0,25	0,33	0,22
Prof. interm	0,19	0,06	0,12	0,00	0,02	0,03	0,06	0,03
Employé	0,19	0,07	0,22	0,02	0,00	0,00	0,01	0,02
Ouvrier qual	0,19	0,06	0,25	0,03	0,00	0,00	0,01	0,02
Ouvrier n-q	0,17	0,06	0,33	0,06	0,01	0,01	0,00	0,03
Inactif	0,10	0,02	0,22	0,03	0,02	0,02	0,03	0,00

- Le dendrogramme correspondant à la CAH :



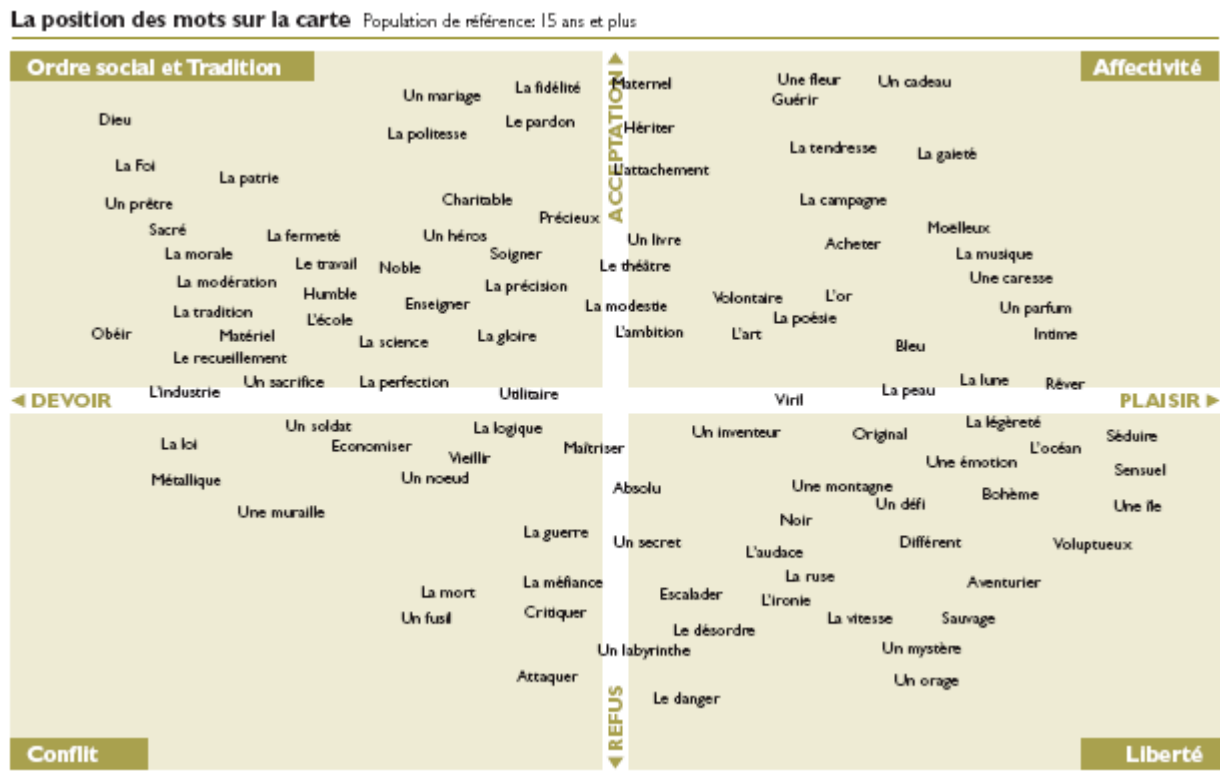
Une CAH analogue, réalisée à partir des individus colonnes, conduit au dendrogramme suivant :



## 5.4.2 Classification à partir d'un tableau Individus x Variables Numériques

Réf. Lebart L., Piron M., Steiner J.-F., La Sémiométrie, Dunod, Paris, 2003.

Dans l'ouvrage cité en référence, les auteurs ont fait le choix de 210 mots. Il est ensuite demandé aux personnes interviewées de noter les mots en fonction de la sensation, agréable ou désagréable, que provoque leur lecture. L'échelle de notation comporte 7 modalités variant de -3 à 3. Pour les traitements statistiques ultérieurs, cette échelle est ramenée à une échelle variant de 1 à 7. L'échantillon interrogé entre 1990 et 2002 s'élève à 11055 personnes. Une enquête analogue, menée pour la Belgique, a conduit au résultat suivant (deux premiers axes d'une ACP) :



On mesure la proximité entre deux mots à l'aide du coefficient de corrélation des séries statistiques obtenues pour les deux mots. Plus précisément, le carré de la distance entre deux mots a et b est égal à  $(1-r(a, b))^2$ , où  $r(a, b)$  désigne le coefficient de corrélation des deux séries. Pour chaque mot, les autres mots qui lui sont le mieux corrélés constituent son champ sémantique interne. Cependant, un même mot peut être corrélé avec des mots non corrélés entre eux.

Une classification ascendante hiérarchique est effectuée à partir de la distance définie précédemment. Il n'est pas évident a priori que des notes fondées seulement sur l'agrément ou le désagrément engendrent des proximités sémantiques. On constate cependant que les classes obtenues regroupent des mots qui ne sont pas de vrais synonymes (la liste de mots excluait a priori la présence de synonymes) mais appartiennent au même halo sémantique. Dans une partition en 12 classes, par exemple, on trouvera rassemblés des mots ayant trait au concept de "sublimation" tels que :

absolu, immense, infini, admirer, adorer, éternel, précieux, secret, sublime.

### Exercice :

A partir de la liste de 7 mots suivants :

efficace, courage, sensuel, montagne, magie, douceur, campagne

imaginez les réponses fournies par dix interviewés et traitez-les à l'aide d'une CAH en utilisant, évidemment, la "distance" 1-r de Pearson.

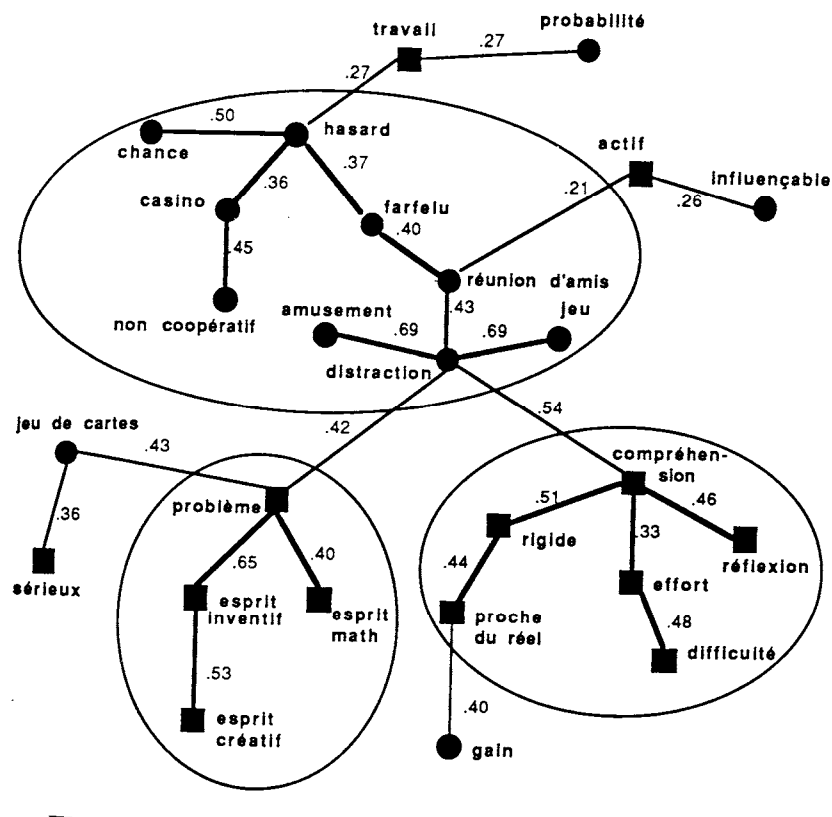
### 5.4.3 Représentation des similitudes par l'arbre de longueur minimale

L'ensemble des  $n$  objets à classer peut être considéré comme un ensemble de points d'un espace. Si l'on ne dispose que des valeurs d'un indice de dissimilarité, on peut représenter les objets par des points (d'un plan par exemple), chaque couple d'objets étant joint par une ligne continue, à laquelle est attachée la valeur de l'indice de dissimilarité. On représente ainsi l'ensemble des objets et des valeurs de l'indice par un graphe complet valué. On cherchera ensuite à extraire de ce graphe un graphe partiel (ayant les mêmes sommets, mais moins d'arêtes) plus aisé à représenter, et permettant néanmoins de bien résumer les valeurs de l'indice. Parmi tous les graphes partiels, ceux qui ont une structure d'arbre sont particulièrement intéressants, car ils peuvent faire l'objet d'une représentation plane. La longueur d'un arbre sera la somme des "longueurs" (valeurs de l'indice) de ses arêtes. Parmi tous les graphes partiels qui sont des arbres, l'arbre de longueur minimale a retenu depuis longtemps l'attention des statisticiens en raison de ses bonnes qualités descriptives, qui ne sont pas étrangères à sa parenté avec les classifications hiérarchiques. On peut, par exemple, montrer l'équivalence avec la classification selon le saut minimal.

Dans la procédure de Kruskal, par exemple, on range les  $n(n - 1)/2$  arêtes dans l'ordre des valeurs croissantes de l'indice. On part des deux premières arêtes, puis on sélectionne successivement toutes les arêtes qui ne font pas de cycle avec les arêtes déjà choisies. On interrompt la procédure dès que l'on a  $n-1$  arêtes. De cette façon, on est sûr d'avoir obtenu un arbre (graphe sans cycle ayant  $n-1$  arêtes).

Exemple : Dans l'ouvrage "Représentations sociales et analyse des données", Doise et al. donnent l'exemple suivant :

Donnons un exemple que Flament emprunte à Abric et Vacherot (1976). Il s'agit d'une recherche effectuée sur la représentation d'une tâche de type «dilemme du prisonnier», tâche qui peut être perçue comme une situation de jeu ou une situation de résolution de problèmes. Les auteurs retiennent 26 termes d'une pré-enquête permettant de traduire l'une ou l'autre de ces situations. Ils demandent ensuite à des sujets ayant effectué une tâche de type dilemme du prisonnier de choisir parmi les 26 termes ceux qui évoquent la situation dans laquelle ils se trouvaient. L'arbre maximum du système de similitude (comprenant 325 corrélations) est de la forme suivante :



Dans cette figure, chaque terme représente un sommet. Le long des liaisons entre sommets (ou arêtes) sont indiqués les indices de similitude. Pour construire un tel arbre, la procédure est la suivante. Il s'agit d'abord d'ordonner les arêtes selon la valeur décroissante de l'indice de similitude qui leur est associé. On retient ensuite les deux premières arêtes qui appartiendront forcément à l'arbre maximum du fait qu'elles ne peuvent être les plus petites dans aucun cycle. Enfin, on ajoute à ces deux premières arêtes, toute arête qui ne forme pas de cycle avec celles déjà retenues. Les arêtes qui sont donc retenues dans l'arbre maximum sont celles qui ne sont minimum dans aucun cycle (voir Degenne et Verges, 1973). Pour illustrer ce propos, prenons l'exemple des éléments Chance, Hasard et Casino qui figurent dans l'arbre maximum ci-dessus. Les arêtes (Chance, Hasard) et (Casino, Hasard) sont inscrites sur le graphe et valent respectivement .50 et .36. On en déduit par conséquent que l'arête (Chance, Casino) est inférieure à .36 ; si tel n'était pas le cas, l'arête (Casino, Hasard) serait supprimée au profit de l'arête (Chance, Casino). En termes de similitude, on peut dire que Chance et Hasard, d'une part, et Hasard et Casino, d'autre part, sont plus proches l'un de l'autre que Chance et Casino.

Sur la base de l'arbre maximum, il est possible de répondre à la question posée par Abric et Vacherot qui est d'identifier les termes associés à jeu ou à résolution de problème comme représentation de la tâche. Flament (1986, 144) en propose la lecture suivante: «Supprimons de l'arbre maximum les arêtes se trouvant entre items de catégories initiales différentes (voir figure) ; les sous-graphes ainsi obtenus sont alors de composition homogène (soit tout jeu, soit tout problème) ; on observe des items isolés (Travail, Probabilité, Actif, etc.), dont la signification initiale est fortement remise en cause (puisque chacun ressemble plus à des items de catégorie opposée qu'aux items de sa propre catégorie). Restent trois sous-graphes importants (indiqués dans la figure) - un pour jeu, deux pour problème -, dont les items voient leur signification initiale confirmée dans la représentation par le voisinage d'items de même catégorie.»