3 Analyse des Correspondances Multiples

3.1 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 à choix multiples et à réponse unique. 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, ..., 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.

Notations utilisées dans ce chapitre :

On note Q l'ensemble des variables (appelées "questions"). On désigne par K_q l'ensemble des modalités de la variable X_q et K l'ensemble de toutes les modalités de réponses.

3.2 Exemple

3.2.1 Enoncé

L'exemple qui suit est extrait de [Crucianu]. Il s'agit d'une partie des données issues de l'enquête "Les étudiants et la ville" effectuée en 2001 par des étudiants de sociologie sous la direction de S. Denèfle à l'Université François Rabelais de Tours.

L'analyse porte sur cinq questions en rapport avec le logement étudiant. L'ensemble des individus statistiques est ici un échantillon de 383 étudiants. Les questions sont les suivantes :

Question	N°	Réponses possibles	Poids (%)	Abréviation
Habitez-vous (variable "mode d'occupation")	1	seul	48,30%	Seul
	2	colocataires	13,84%	Coloc
	3	en couple	13,05%	Couple
	4	avec les parents	23,50%	Parents
	5	non réponse	1,31%	NR1
Quel type d'habitation occupez-vous ? (variable "type d'habitation")	6	cité universitaire	10,70%	Cité
	7	studio	28,20%	Studio
	8	appartement	30,29%	Appart
	9	chambre chez un particulier	5,22%	Chambre
	10	autre	19,84%	Autre
	11	non réponse	5,74%	NR2
Si vous vivez en dehors du foyer familial, depuis combien de temps ? (variable "ancienneté")	12	moins de 1 an	20,89%	< 1 an
	13	1 à 3 ans	24,80%	1-3 ans
	14	plus de 3 ans	28,72%	> 3 ans
	15	non applicable	24,80%	NA
	16	non réponse	0,78%	NR3
A quelle distance approximative de la Fac	17	moins de 1 km	26,89%	< 1 km
	18	1 à 5 km	49,87%	1 à 5 km

vivez-vous ? (variable "éloignement")	19	plus de 5 km	20,89%	> 5 km
	20	non réponse	2,35%	NR4
Quelle est la superficie de votre logement ? (variable "superficie")	21	moins de 10 m^2	9,14%	< 10 m^2
	22	10 à 20 m^2	17,75%	10 à 20 m^2
	23	20 à 30 m^2	24,80%	20 à 30 m^2
	24	plus de 30 m^2	39,16%	> 30 m^2
	25	non réponse	9,14%	NR5

3.2.2 Différentes représentations des données recueillies

Nous verrons ultérieurement qu'il est préférable de regrouper les modalités dont la fréquence est trop faible (inférieure à 5% par exemple) avec d'autres modalités. Aussi, dans les données qui suivent, les modalités "Parents" et "NR1" ont été regroupées pour la variable "mode", de même que "NA" et "NR3" pour la variable "ancienneté" et ">5km" et "NR4" pour la variable "éloignement". Il reste donc 22 modalités distinctes.

Les données recueillies peuvent être représentées, de façon classique, à l'aide d'un tableau protocole ou d'un tableau d'effectifs. Cependant, deux autres représentations sont également utilisées : le tableau disjonctif complet (TDC) et le tableau de Burt (TdB).

3.2.2.1 Tableau protocole et tableau d'effectifs

Les données recueillies peuvent être représentées, de façon classique, à l'aide d'un tableau protocole ou d'un tableau d'effectifs :

Tableau protocole

Sujet	Mode d'occupation	Type d'habitation	Ancienneté	Eloignement	Superficie
S1	seul	cité	< 1 an	< 1 km	< 10 m^2
S2	coloc	appart	1 à 3 ans	1 à 5 km	20 à 30 m^2
...	...				

Tableau d'effectifs

Mode d'occupation	Type d'habitation	Ancienneté	Eloignement	Superficie	Effectif
seul	cité	< 1 an	< 1 km	< 10 m^2	7
seul	cité	< 1 an	< 1 km	10 à 20 m^2	2
...					

3.2.2.2 Tableau disjonctif complet (TDC)

Le tableau disjonctif complet 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 correspondante ou non :

	Mode d'occupation				Type habitation						Ancienneté				Eloignement			Superficie				
	Seul	Coloc	Couple	Parents et NR1	Cité	Studio	Appart	Chambre	Autre	NR2	<= 1 an	1-3 ans	> 3 ans	NA et NR3	- de 1km	1 à 5 km	+ 5 km et NR4	- de 10 m^2	10 à 20 m^2	20 à 30 m^2	+ de 30 m^2	NR5
i1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0
i2	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0
i3	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0

i4	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1
i5	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0
...	...	...																				

3.2.2.3 Tableau disjonctif des patrons

Un patron de réponse, c'est une combinaison de modalités susceptible d'être choisie par un sujet. Ici, le nombre de patrons possible est très élevé : $4 \times 6 \times 4 \times 3 \times 5 = 1440$. Autrement dit, la plupart d'entre eux ne sont pas présents dans les réponses observées.

En regroupant les lignes identiques dans le tableau disjonctif complet ou en convertissant en tableau disjonctif le tableau d'effectifs, on obtient le tableau disjonctif des patrons de réponses. Par exemple :

	Mode d'occupation				Type habitation						Ancienneté				Eloignement			Superficie				
	Seul	Coloc	Couple	Parents et NR 1	Cité	Studio	Appart	Chambre	Autre	NR 2	<= 1 an	1-3 ans	> 3 ans	NA et NR 3	- de 1k m	1 à 5 km	+ 5 km et NR 4	- de 10 m^2	10 à 20 m^2	20 à 30 m^2	+ de 30 m^2	NR 5
p1	12	0	0	0	12	0	0	0	0	0	12	0	0	0	12	0	0	12	0	0	0	0
p2	6	0	0	0	0	6	0	0	0	0	6	0	0	0	0	6	0	0	6	0	0	0
...	...	...																				

3.2.2.4 Tableau de Burt (TdB)

Le tableau de Burt 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. Pour l'exemple étudié, le tableau de Burt est le suivant :

	Seul	Coloc	Couple	Parents & NR	Cité	Studio	Appart	Chambre	Autre	NR 2	<= 1 an	1-3 ans	> 3 ans	NA & NR	- de 1k m	1 à 5 km	+5 km & NR	- de 10 m^2	10 à 20 m^2	20 à 30 m^2	+ de 30 m^2	NR 5
Seul	185	0	0	0	34	90	40	13	3	5	61	60	59	5	70	101	14	32	61	71	21	0
Colo	0	53	0	0	5	6	32	2	3	5	13	18	21	1	13	33	7	1	4	8	40	0
Coup	0	0	50	0	2	10	34	0	3	1	5	14	28	3	15	23	12	2	2	14	32	0
Par / NR	0	0	0	95	0	2	10	5	67	11	1	3	2	89	5	34	56	0	1	2	57	35
Cité	34	5	2	0	41	0	0	0	0	0	17	13	9	2	15	23	3	27	9	1	4	0
Stud	90	6	10	2	0	108	0	0	0	0	29	33	45	1	41	61	6	1	33	57	17	0
App	40	32	34	10	0	0	116	0	0	0	23	35	47	11	37	62	17	1	10	29	74	2
Cha	13	2	0	5	0	0	0	20	0	0	6	6	3	5	6	10	4	4	7	5	4	0
Autr	3	3	3	67	0	0	0	0	76	0	2	4	4	66	2	29	45	0	1	1	50	24
NR2	5	5	1	11	0	0	0	0	0	22	3	4	2	13	2	6	14	2	8	2	1	9
- de1	61	13	5	1	17	29	23	6	2	3	80	0	0	0	30	44	6	14	26	24	16	0
1-3	60	18	14	3	13	33	35	6	4	4	0	95	0	0	25	60	10	11	22	28	32	2
+de3	59	21	28	2	9	45	47	3	4	2	0	0	110	0	43	53	14	10	14	41	45	0
NA / NR	5	1	3	89	2	1	11	5	66	13	0	0	0	98	5	34	59	0	6	2	57	33
- 1k	70	13	15	5	15	41	37	6	2	2	30	25	43	5	103	0	0	12	26	38	26	1
1 à 5	101	33	23	34	23	61	62	10	29	6	44	60	53	34	0	191	0	20	35	47	82	7
+ 5k/NR	14	7	12	56	3	6	17	4	45	14	6	10	14	59	0	0	89	3	7	10	42	27
- 10	32	1	2	0	27	1	1	4	0	2	14	11	10	0	12	20	3	35	0	0	0	0
10-20	61	4	2	1	9	33	10	7	1	8	26	22	14	6	26	35	7	0	68	0	0	0
20-30	71	8	14	2	1	57	29	5	1	2	24	28	41	2	38	47	10	0	0	95	0	0
30+	21	40	32	57	4	17	74	4	50	1	16	32	45	57	26	82	42	0	0	0	150	0
NR5	0	0	0	35	0	0	2	0	24	9	0	2	0	33	1	7	27	0	0	0	0	35

Lecture de ce tableau :

- parmi les 383 étudiants interrogés, 185 logent seuls ;
- parmi les 383 étudiants interrogés, 34 logent seuls, en cité universitaire ;
- etc.

Le tableau de Burt possède de nombreuses propriétés remarquables :

- Le tableau est symétrique : $n_{ij} = n_{ji}$;
- Les encadrés situés le long de la diagonale principale (du haut à gauche vers le bas à droite) donnent les effectifs correspondant à chaque modalité ;
- Les autres encadrés sont les tableaux de contingence correspondant aux variables prises deux à deux ;
- La ligne j (ou la colonne j) du tableau est la somme des lignes du TDC correspondant aux individus qui possèdent la modalité j ;
- La somme des nombres situés sur une même ligne est égale au terme diagonal de la ligne multiplié par le nombre de variables ; propriété identique pour les colonnes ;
- La somme des nombres situés dans un encadré est égal à l'effectif total ;
- La somme de tous les nombres du tableau est égale à l'effectif total multiplié par le carré du nombre de variables.

Le tableau de Burt peut être vu comme une juxtaposition de tableaux de contingence. Il peut être obtenu facilement à partir du tableau disjonctif complet. En revanche, il n'existe pas de moyen simple permettant de recomposer le tableau disjonctif complet (ou l'un des autres tableaux équivalents) à partir du tableau de Burt. De plus, plusieurs protocoles différents peuvent conduire au même tableau de Burt.

3.2.3 Effectifs théoriques et taux de liaison calculés à partir du tableau de Burt

Le tableau de Burt est composé de sous-tableaux diagonaux (effectifs des différentes questions) et non diagonaux (tableaux de contingence des variables prises deux à deux). Le calcul du tableau des effectifs théoriques correspondants conduit aux résultats suivants :

- pour un sous-tableau non diagonal, les effectifs théoriques sont les mêmes que ceux du tableau de contingence correspondant ;
- pour un sous-tableau diagonal, l'effectif théorique de la cellule correspondant au croisement de deux modalités est égal au produit de l'effectif de l'une par la fréquence de l'autre.

Pour les sous-tableaux non diagonaux, les taux de liaison sont ceux du tableau de contingence correspondant.

Pour un sous-tableau diagonal :

- les cellules non diagonales ont un taux de liaison égal à -1 ;
- le taux de liaison de la cellule diagonale correspondant à la modalité *k* est égal à :

$$t_{kk} = \frac{Nombre\ d'\ individus\ n'\ ayant\ pas\ choisi\ la\ modalité\ k}{Nombre\ d'\ individus\ ayant\ choisi\ la\ modalité\ k} = \frac{1 - f_k}{f_k}$$

Ce rapport pourrait aussi être appelé l'inverse de la cote de la modalité *k*.

3.3 Distances entre individus, entre modalités. Inertie du nuage

L'ACM peut être considérée comme une variante de l'AFC. 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).

Nous pouvons donc, comme en AFC, nous intéresser aux profils ligne et colonne, aux taux de liaison et au Φ^2 du tableau disjonctif complet, vu comme un tableau de contingence. Nous avons vu que la métrique du Φ^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.

3.3.1 Profils lignes et colonnes moyens pour le tableau disjonctif complet et le tableau de Burt

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

3.3.1.1 Profils colonnes moyens

Dans le tableau disjonctif complet, chaque ligne représente un individu statistique, avec la fréquence 1/N. Le profil colonne moyen attribue donc la fréquence 1/N à chaque ligne du tableau.

Le profil colonne moyen du tableau disjonctif des patrons est formé des fréquences des différents patrons de réponses dans la population étudiée.

Dans le tableau de Burt, le profil colonne moyen est identique au profil ligne moyen (étudié dans le paragraphe suivant), car le tableau est symétrique.

3.3.1.2 Profils lignes moyens

D'après la propriété d'équivalence distributionnelle, le profil ligne moyen du tableau disjonctif complet et celui du tableau disjonctif des patrons sont identiques. En effet, on passe du premier au second en regroupant des lignes identiques.

Le profil ligne moyen du tableau disjonctif complet est obtenu en calculant, pour chaque modalité, le quotient de sa fréquence par le nombre Q de questions.

En notant respectivement n_k et f_k l'effectif et la fréquence de la modalité k, on a :

$$f_k = \frac{n_k}{N} = \frac{Nombre\ \ d'individus\ \ ayant\ \ choisi\ \ la\ \ modalité\ \ k}{Nombre\ \ total\ \ d'individus}$$

et le k-ième élément du profil-ligne moyen est :

$$f'_k = \frac{f_k}{Q} = \frac{n_k}{QN} = \frac{Nombre\ d'individus\ ayant\ choisi\ la\ modalité\ k}{Nombre\ de\ questions \times Nombre\ total\ d'individus}$$

Ainsi, dans notre exemple, la fréquence de la modalité "Seul" de la variable "Mode d'occupation" est 0,483, alors que le nombre de questions est Q=5. La première valeur du profil ligne moyen est donc :

$$\frac{0,483}{5} = 0,0966.$$

N.B. Dans ce chapitre, f_k et f'_k désignent des quantités différentes : f_k est la fréquence de la modalité k dans la population étudiée; f'_k est définie comme pour l'AFC, fréquence ligne marginale de la k-ième colonne du tableau disjonctif des patrons.

Pour le tableau de Burt, le profil ligne moyen et le profil colonne moyen sont identiques, car le tableau est symétrique. On peut montrer que le k-ième élément du profil ligne moyen est donné par:

$$\frac{Nombre\ de\ questions \times Nombre\ d'individus\ ayant\ choisi\ la\ modalité\ k}{Nombre\ de\ questions \times Nombre\ de\ questions \times Nombre\ total\ d'individus} = \frac{Q\,n_k}{Q^2 N} = f'_k$$

On retrouve donc le même profil ligne moyen que pour les deux autres tableaux.

3.3.2 Distances entre individus (profils lignes du tableau disjonctif des patrons)

Remarque. Dans notre exemple, les données effectivement observées nous sont données sous forme de tableau de Burt. Il n'est donc pas possible de représenter de manière exacte les distances entre individus (ni même de savoir exactement quels sont les patrons de réponses effectivement observés). Cependant, il est possible, à partir d'un tableau de Burt, de générer l'un des tableaux protocoles possibles conduisant à ce tableau de Burt.
C'est ce qui a été fait ici (cf. feuille de données Etudiants-ville-2006-TDP du classeur Etudiants-ville-2010.stw). Ce tableau "calculé" comporte 142 patrons différents (nombre sans doute plus élevé que ce qui a été réellement observé). Les 18 patrons d'effectif supérieur ou égal à 5 sont les suivants :

	Effectif	Seul	Colloc	Couple	Parents & NR	Cite	Studio	Appart	Chambre	Autre	NR2	< 1 an	1-3 ans	> 3 ans	NA & NR	< 1 km	1-5 km	> 5 km & NR	< 10 m2	10 a 20 m2	20 a 30 m2	> 30 m2	NR5
Parents/Autre/NA/>5km/>30m	22	0	0	0	22	0	0	0	0	22	0	0	0	0	22	0	0	22	0	0	0	22	0
Parents/Autre/NA/>5km/NR5	20	0	0	0	20	0	0	0	0	20	0	0	0	0	20	0	0	20	0	0	0	0	20
Parents/Autre/NA/1-5km/>30m	19	0	0	0	19	0	0	0	0	19	0	0	0	0	19	0	19	0	0	0	0	19	0
Seul/Studio/>3a/<1km/20a30m	15	15	0	0	0	0	15	0	0	0	0	0	0	15	0	15	0	0	0	0	15	0	0
Seul/Studio/1-3a/1-5km/20-30m	12	12	0	0	0	0	12	0	0	0	0	0	12	0	0	0	12	0	0	0	12	0	0
Seul/Studio/>3a/1-5km/20-30m	11	11	0	0	0	0	11	0	0	0	0	0	0	11	0	0	11	0	0	0	11	0	0
Coloc/Appart/1-3a/1a5km/>30m	10	0	10	0	0	0	0	10	0	0	0	0	10	0	0	0	10	0	0	0	0	10	0
Seul/Studio/<1an/<1km/10-20m	9	9	0	0	0	0	9	0	0	0	0	9	0	0	0	9	0	0	0	9	0	0	0
Seul/Cite/1-3an/1-5km/<10m	8	8	0	0	0	8	0	0	0	0	0	0	8	0	0	0	8	0	8	0	0	0	0
Seul/Studio/<1a/1-5km/10-20m	8	8	0	0	0	0	8	0	0	0	0	8	0	0	0	0	8	0	0	8	0	0	0
Coloc/Appart/>3a/1-5km/>30m	8	0	8	0	0	0	0	8	0	0	0	0	0	8	0	0	8	0	0	0	0	8	0
Couple/Appart/>3a/<1km/>30m	8	0	0	8	0	0	0	8	0	0	0	0	0	8	0	8	0	0	0	0	0	8	0
Seul/Cite/<1a/1-5km/<10m	7	7	0	0	0	7	0	0	0	0	0	7	0	0	0	0	7	0	7	0	0	0	0
Seul/Studio/<1a/1-5km/20-30m	7	7	0	0	0	0	7	0	0	0	0	7	0	0	0	0	7	0	0	0	7	0	0
Seul/Studio/1-3a/1-5km/10-20m	7	7	0	0	0	0	7	0	0	0	0	0	7	0	0	0	7	0	0	7	0	0	0
Seul/Appart/1-3a/<1km/20a30m	6	6	0	0	0	0	0	6	0	0	0	0	6	0	0	6	0	0	0	0	6	0	0
Parents/NR2/NA/>5km/NR5	6	0	0	0	6	0	0	0	0	0	6	0	0	0	6	0	0	6	0	0	0	0	6
Seul/Cite/>3a/<1km/<10m	5	5	0	0	0	5	0	0	0	0	0	0	0	5	0	5	0	0	5	0	0	0	0

Comme en AFC, la distance utilisée est la *métrique du* Φ^2, appliquée au tableau disjonctif des patrons de réponses, considéré comme tableau de contingence. Cependant, compte tenu de la structure particulière du tableau de contingence utilisé, les distances entre individus lignes, et la distance d'un individu ligne au profil moyen prennent les significations suivantes.

3.3.2.1 Distance d'un individu ou d'un patron au profil colonne moyen

Tous les individus qui ont choisi le même patron de réponse sont représentés par un même point de l'espace multidimensionnel. La distance d'un individu au profil moyen est égale à la distance de son patron de réponse au profil colonne moyen.

On peut montrer que la distance d'un patron au profil ligne moyen ne dépend que de la fréquence des modalités qui le composent, et du nombre de questions. Plus précisément, la distance d'un patron au profil ligne moyen est :

$$d^2_{\Phi^2}(O,\ L_i) = \left(\frac{1}{Q}\sum_k \frac{\delta_{ik}}{f_k}\right) - 1$$

où δ_{ik} vaut 1 si la modalité k fait partie du patron i et 0 sinon.

Autrement dit, un patron sera d'autant plus loin de l'origine qu'il fait intervenir des modalités plus rares. On peut aussi écrire cette formule sous la forme :

$$d^2_{\Phi^2}(O,\ Patron\ \ i) = \left(\frac{1}{Nombre\ \ de\ \ Questions} \sum \frac{1}{fréquence\ \ de\ \ la\ \ modalité\ \ k} \right) - 1$$

où la somme est étendue à toutes les modalités faisant partie du patron i.

Exemple :

Le premier patron cité ci-dessus est formé des modalités Parents, Autre, NA & NR, > 5 kms, > 30 m2, dont les effectifs respectifs sont : 95, 76, 98, 89 et 150.

Sa distance au profil moyen est donnée par :

$$d^2_{\Phi^2}(O, Patron1) = \frac{1}{5}\left(\frac{383}{95} + \frac{383}{76} + \frac{383}{98} + \frac{383}{89} + \frac{383}{150} \right) - 1 = 2{,}97$$

3.3.2.2 Distance entre deux patrons ou entre deux individus

De même, la distance entre deux patrons, ou de deux individus ayant choisi ces patrons, dépend du nombre de questions et de la fréquence des modalités qui appartiennent à l'un ou l'autre des deux patrons sans appartenir aux deux simultanément. Plus précisément, la formule est la suivante :

$$d^2_{\Phi^2}(Patron\ \ i, Patron\ \ i') = \frac{1}{Nb\ \ de\ \ Questions} \sum \frac{1}{fréquence\ \ de\ \ la\ \ modalité\ \ k}$$

où la somme est étendue à toutes les modalités faisant partie de l'un des deux patrons, sans faire partie des deux patrons.

Exemple :

Les deux premiers patrons cités ci-dessus ne diffèrent que par leur modalité sur la question 5 : "> 30 m2", d'effectif 150 pour le premier, "NR5", d'effectif 35 pour le second. La distance entre ces deux patrons est donnée par :

$$d^2_{\Phi^2}(Patron1, Patron2) = \frac{1}{5}\left(\frac{383}{150} + \frac{383}{35} \right) = 2{,}70$$

3.3.3 Distances entre modalités (profils colonnes du tableau disjonctif)

3.3.3.1 Distance d'une modalité de réponse au profil colonne moyen

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

$$d^2_{\Phi^2}(O,\ M_k) = \frac{1}{f_k} - 1 = \frac{n}{n_k} - 1 = \frac{Effectif\ \ total}{Effectif\ \ de\ \ k} - 1$$

On retrouve ainsi le taux de liaison correspondant à la modalité.

Pour la modalité "Seul", on obtient, par exemple :

$$d^2_{\Phi^2}(O,\ Seul) = \frac{383}{185} - 1 = 1{,}07$$

La formule montre qu'une modalité sera d'autant plus loin du profil moyen que sa fréquence est faible. Par exemple, la modalité "Chambre" vérifie : d^2=16,41 (c'est celle dont l'effectif est le plus faible), alors que la modalité "de 1 à 5 km" vérifie : d^2=1,005 (c'est celle dont l'effectif est le plus élevé). Il est difficile

de se reporter aux graphiques qui suivent pour visualiser ces distances en raison des déformations dues aux projections.

3.3.3.2 Distance entre deux modalités de réponse

On peut montrer que la distance entre les modalités k et k' est donnée par :

$$d^2_{\Phi^2}(M_k,\ M_{k'}) = \frac{1}{f_k} + \frac{1}{f_{k'}} - 2\frac{f_{kk'}}{f_k f_{k'}} = \frac{n_k + n_{k'} - 2n_{kk'}}{n_k n_{k'}/n}$$

où $f_{kk'}$ est la fréquence de la combinaison de modalités k et k', ou encore :

$$d^2_{\Phi^2}(M_k,\ M_{k'}) = \frac{Effectif\ \ de\ \ k + Effectif\ \ de\ \ k' - 2 \times Effectif\ \ de\ \ la\ \ combinaison\ \ k\ \&\ k'}{Effectif\ \ de\ \ k \times\ Effectif\ \ de\ \ k' / Effectif\ \ total}$$

Par exemple, sachant que l'effectif de la modalité "Seul" est 185, celui de la modalité "Cité" est 41, celui de la combinaison "Seul et en cité" est 34 et l'effectif total est de 383, on obtient :

$$d^2_{\Phi^2}(Seul,\ Cité) = \frac{185 + 41 - 2 \times 34}{185 \times 41/383} = 7{,}978$$

La formule montre que la distance sera d'autant plus faible que l'effectif conjoint est proche des effectifs de chacune des deux modalités, particulièrement si celles-ci sont d'effectifs élevés. Par exemple, pour les modalités "Seul" et "de 1 à 5 km", dont les effectifs sont respectivement 185 et 191, alors que l'effectif conjoint est de 101, on obtient : d^2=1,89.

En particulier, l'indépendance des modalités *k* et *k'* se traduit par :

$$d^2_{\Phi^2}(M_k, M_{k'}) = d^2_{\Phi^2}(O, M_k) + d^2_{\Phi^2}(O, M_{k'})$$

Ainsi, dans l'espace multidimensionnel, deux modalités indépendantes sont représentées par des points M_k et $M_{k'}$ tels que le triangle $OM_kM_{k'}$ soit un triangle rectangle en O. Si les modalités *k* et *k'* s'attirent, l'angle (OM_k, $OM_{k'}$) est un angle aigu, si elles se repoussent, (OM_k, $OM_{k'}$) est un angle obtus.

3.4 Inertie du nuage de points. Contributions

Dans le cas de l'ACM, les coefficients Φ^2 et χ^2 ne sont pas véritablement informatifs. En effet, pour le tableau disjonctif complet, ou le tableau disjonctif des patrons, considérés comme des tableaux de contingence, le coefficient Φ^2 vaut :

$$\Phi^2 = \frac{K - Q}{Q} = \frac{Nombre\ de\ modalités - Nombre\ de\ questions}{Nombre\ de\ questions}$$

où K désigne le nombre de modalités et Q le nombre de questions. Cette quantité représente aussi l'inertie du nuage des individus ou du nuage des modalités.

Dans notre exemple, on a : K=22, Q=5, et donc : $\Phi^2 = \frac{22-5}{5} = 3{,}4$.

La contribution absolue d'une modalité à l'inertie du nuage de points est :

$$Cta(M_k) = \frac{1 - f_k}{Q} = \frac{1 - fréquence\ \ de\ \ k}{Nombre\ \ de\ \ questions}$$

La contribution relative de cette modalité à l'inertie du nuage de points est :

$$Ctr(M_k) = \frac{1 - f_k}{K - Q} = \frac{1 - \textit{fréquence de k}}{\textit{Nombre de modalités} - \textit{Nombre de questions}}$$

Par exemple, pour la modalité "Seul" :

$$Cta(Seul) = \frac{1 - 0,483}{5} = 0,1034$$

Sa contribution relative est obtenue en divisant par l'inertie totale du nuage (3,4 dans notre exemple) :

$$Ctr(Seul) = \frac{0,1034}{3,4} = 0,0304$$

L'inertie totale peut être exprimée comme la somme des inerties de chacune des variables. Mais l'inertie de la variable X_q est donnée par : $I(X_q) = \frac{K_q - 1}{Q}$, où K_q est le nombre de modalités de la variable X_q. Par exemple, pour la première variable :

$$I(X_1) = \frac{4 - 1}{5} = 0,6$$

L'inertie relative d'une variable est obtenue en divisant son inertie absolue par celle du nuage, c'est-à-dire par Φ^2. Compte tenu des formules précédentes, on a encore :

$$Inr(X_q) = \frac{K_q - 1}{K - Q} = \frac{(\textit{Nombre de modalités de la question q}) - 1}{\textit{Nombre de Modalités} - \textit{Nombre de Questions}}$$

Autrement dit, l'influence d'une variable dépend seulement du nombre de ses modalités. Pour éviter que certaines variables prennent une importance excessive, ou au contraire soient peu présentes dans l'analyse, il faut donc éviter des différences trop marquées entre les nombres de modalités des variables à analyser.

Par exemple, pour la première variable :

$$Inr(X_1) = \frac{0,6}{3,4} = \frac{4 - 1}{22 - 5} = 0,1765$$

Remarque : Pour le tableau de Burt, considéré comme tableau de contingence, les coefficients Φ^2 et χ^2 sont différents de ceux du tableau disjonctif complet. Mais le module ACM de Statistica, par exemple, indique en fait la valeur du Φ^2 correspondant au tableau disjonctif complet.

3.5 L'analyse des correspondances multiples proprement dite

L'analyse peut être menée à partir du tableau disjonctif complet (ou de l'un des tableaux qui lui sont équivalents) ou du tableau de Burt. Les deux méthodes conduisent à des résultats analogues (mais pas identiques).

D'un point de vue mathématique, le traitement opéré sur les données du tableau de Burt est identique à celui opéré sur un tableau de contingence lors d'une AFC. On obtiendra donc, comme lors d'une AFC :

- Des axes factoriels associés à des valeurs propres ;
- Pour chaque ligne (ou colonne) du tableau de Burt, des coordonnées, des contributions à la formation des axes et des qualités de représentation.

Cependant, la méthode diffère de l'AFC par divers aspects, et nous devons adapter notre grille d'interprétation.

Le tableau de Burt étant symétrique, les profils lignes et les profils colonnes sont identiques. Il en est de même des coordonnées des individus-lignes et des individus-colonnes. Nous nous intéresserons donc uniquement aux individus-colonnes (par exemple).

Les profils-lignes du tableau de Burt (qui sont ici identiques aux profils-colonnes) ne sont pas directement interprétables. Le nuage des modalités a cependant une propriété intéressante : le centre de gravité des différentes modalités d'une même question est l'origine des axes.

3.5.1 Valeurs propres

L'analyse du tableau de Burt produit au plus *K*-*Q* valeurs propres non nulles. La décroissance de ces valeurs propres est beaucoup plus lente que dans le cas de l'AFC. Benzecri (1992) a élaboré une méthode permettant de calculer des "taux modifiés" d'inertie expliquée par chaque valeur propre.

Nombre de Dims.	Valeurs Propres et Inertie de toutes les Dimensions (Etudiants-ville-2006.sta)				
	Inertie Totale = 3,40				
	ValSing.	ValProp.	%age Inertie	%age Cumulé	Chi^2
1	0,8494	0,7215	21,22	21,22	2306,12
2	0,6351	0,4033	11,86	33,08	1288,98
3	0,5734	0,3288	9,67	42,75	1051,04
4	0,5019	0,2519	7,41	50,16	805,13
5	0,4859	0,2361	6,94	57,11	754,60
6	0,4588	0,2105	6,19	63,30	672,78
7	0,4455	0,1985	5,84	69,14	634,44
8	0,4318	0,1865	5,48	74,62	596,03
9	0,4133	0,1708	5,02	79,65	545,91
10	0,4084	0,1668	4,90	84,55	533,01
11	0,3521	0,1240	3,65	88,20	396,27
12	0,3323	0,1104	3,25	91,44	352,93
13	0,3229	0,1042	3,07	94,51	333,20
14	0,2733	0,0747	2,20	96,71	238,73
15	0,2352	0,0553	1,63	98,33	176,81
16	0,1961	0,0385	1,13	99,47	122,90
17	0,1348	0,0182	0,53	100,00	58,10

3.5.1.1 Taux modifiés proposés par Benzécri

La transformation proposée par Benzécri est basée sur la comparaison des valeurs propres obtenues d'une part par l'AFC, d'autre part par l'ACM, pour une situation à deux variables.

La somme des valeurs propres est égale à l'inertie totale, c'est-à-dire $\frac{K-Q}{Q}$ et la moyenne des valeurs propres est égale à $\lambda_m = \frac{1}{Q} = \frac{1}{Nb\ \ de\ \ questions}$. On ne conserve que les valeurs propres λ supérieures à λ_m et on calcule pour chacune d'entre elles : $\lambda' = (\lambda - \lambda_m)^2$. Le taux d'inertie modifié est alors calculé par : $\frac{\lambda'}{\sum \lambda'}$ et on conserve les valeurs propres dont le taux modifié est supérieur à la moyenne (des taux modifiés). Pour l'exemple traité, l'application de cette méthode donne les résultats suivants :

La moyenne des valeurs propres est : $\lambda_m = \frac{1}{5} = 0{,}2$, ce qui conduit à ne conserver que les 6 premières valeurs propres. La transformation précédente donne alors :

Nb de dim.	Val Prop.	$\lambda' = (\lambda - \lambda_m)^2$	Taux d'inertie modifié
1	0,7215	0,2720	81,43%

2	0,4033	0,0413	12,37%
3	0,3288	0,0166	4,97%
4	0,2519	0,0027	0,81%
5	0,2361	0,0013	0,39%
6	0,2105	0,0001	0,03%

Le taux d'inertie modifié moyen est de 100%/6= 16,7%. Seule la première valeur propre dépasse ce taux, mais une étude limitée seulement au premier axe principal présenterait peu d'intérêt. Nous étudierons donc les deux premiers axes (voire, éventuellement, le 3ème).

Remarque : Selon Benzécri, les taux modifiés représentent l'écart du nuage de points par rapport au nuage parfaitement sphérique qui serait obtenu si aucun lien n'existait entre les modalités.

3.5.2 Résultats relatifs aux modalités

Le tableau ci-dessous donne les coordonnées, la contribution (inertie relative) et la qualité de représentation (cos^2) de chacune des modalités selon les trois premiers axes principaux. Il donne également le poids de chaque modalité et sa contribution à la formation de l'inertie totale.

	Ligne	Coord. dim 1	Coord. dim 2	Coord. dim 3	Masse	Qualité	Inertie	Inertie dim 1	Cos^2 dim 1	Inertie dim 2	Cos^2 dim 2	Inertie dim 3	Cos^2 dim 3
MODE:Seul	1	-0,6921	0,5251	-0,2405	0,0966	0,7592	0,0304	0,0641	0,4475	0,0661	0,2576	0,0170	0,0540
MODE:Coloc	2	-0,2275	-1,0201	0,7869	0,0277	0,2749	0,0507	0,0020	0,0083	0,0714	0,1671	0,0521	0,0995
MODE:Couple	3	-0,2219	-1,3155	0,2448	0,0261	0,2762	0,0511	0,0018	0,0074	0,1120	0,2598	0,0048	0,0090
MODE:Parents et NR	4	1,5914	0,2388	-0,0995	0,0496	0,8575	0,0442	0,1741	0,8354	0,0070	0,0188	0,0015	0,0033
TYPE:Cité	5	-0,7505	1,4395	1,9098	0,0214	0,7532	0,0525	0,0167	0,0675	0,1100	0,2484	0,2375	0,4373
TYPE:Studio	6	-0,7176	0,1556	-1,0249	0,0564	0,6243	0,0422	0,0403	0,2022	0,0034	0,0095	0,1802	0,4125
TYPE:Appart	7	-0,2390	-1,0433	0,3581	0,0606	0,5534	0,0410	0,0048	0,0248	0,1635	0,4729	0,0236	0,0557
TYPE:Chambre	8	-0,2358	0,7976	0,0280	0,0104	0,0382	0,0558	0,0008	0,0031	0,0165	0,0351	0,0000	0,0000
TYPE:Autre	9	1,5850	0,1322	-0,0137	0,0397	0,6263	0,0472	0,1382	0,6219	0,0017	0,0043	0,0000	0,0000
TYPE:NR2	10	0,9207	0,8725	-0,3942	0,0115	0,1075	0,0554	0,0135	0,0517	0,0217	0,0464	0,0054	0,0095
ANC:< 1 an	11	-0,6570	0,5958	0,1375	0,0418	0,2127	0,0465	0,0250	0,1140	0,0368	0,0937	0,0024	0,0050
ANC:1-3 ans	12	-0,4743	-0,0604	0,1251	0,0496	0,0806	0,0442	0,0155	0,0742	0,0004	0,0012	0,0024	0,0052
ANC:> 3ans	13	-0,4839	-0,6110	-0,1348	0,0574	0,2521	0,0419	0,0186	0,0944	0,0532	0,1504	0,0032	0,0073
ANC:NA et NR	14	1,5393	0,2579	-0,0823	0,0512	0,8400	0,0438	0,1681	0,8148	0,0084	0,0229	0,0011	0,0023
ELOIGN:< 1km	15	-0,6523	0,0477	-0,2336	0,0538	0,1774	0,0430	0,0317	0,1565	0,0003	0,0008	0,0089	0,0201
ELOIGN:1 à 5 km	16	-0,2129	-0,0935	0,1793	0,0997	0,0858	0,0295	0,0063	0,0451	0,0022	0,0087	0,0097	0,0320
ELOIGN:>5 km - NR	17	1,2118	0,1454	-0,1143	0,0465	0,4549	0,0452	0,0946	0,4445	0,0024	0,0064	0,0018	0,0040
SUP:< 10 m^2	18	-0,7762	1,5959	2,0488	0,0183	0,7389	0,0534	0,0153	0,0606	0,1154	0,2561	0,2333	0,4222
SUP:10 à 20 m^2	19	-0,6118	0,7532	-0,5435	0,0355	0,2670	0,0484	0,0184	0,0808	0,0499	0,1225	0,0319	0,0638
SUP:20 à 30 m^2	20	-0,6689	-0,1651	-0,8958	0,0496	0,4213	0,0442	0,0308	0,1476	0,0034	0,0090	0,1210	0,2647
SUP:> 30 m^2	21	0,4169	-0,7995	0,4513	0,0783	0,6545	0,0358	0,0189	0,1119	0,1241	0,4115	0,0485	0,1311
SUP:NR5	22	1,9938	0,8154	-0,4956	0,0183	0,4914	0,0534	0,1007	0,3998	0,0301	0,0669	0,0136	0,0247

On retrouve dans ce tableau l'inertie relative de chaque variable, comme somme des inerties relatives des modalités qui la compose. Par exemple, pour la première variable :

$$\frac{I(X_1)}{I} = \frac{0,6}{3,4} = 0,0304 + 0,0507 + 0,0511 + 0,0442 = 0,1765$$

L'interprétation utilisera essentiellement les modalités qui ont les meilleures qualités de représentation selon chacun des axes factoriels (colonnes Cosinus^2) ou dans l'espace factoriel retenu (colonne "Qualité"). Mais, il faudra retenir les modalités jusqu'à un seuil assez bas, 0,25 par exemple.

Enfin, l'inertie relative par rapport à chaque axe permettra de retenir les modalités qui ont le plus fortement contribué à la formation de cet axe. On pourra par exemple, retenir les modalités dont l'inertie relative par rapport à un axe dépasse 1/22 c'est-à-dire 0,045.

3.5.3 Résultats graphiques et interprétation

L'interprétation des résultats d'une ACM est souvent assez délicate, en raison de la faible décroissance des valeurs propres, et du grand nombre de modalités, ce qui rend les graphiques assez peu lisibles.

Selon Benzécri, interpréter un axe consiste à trouver ce qui est similaire d'une part entre tous les éléments figurant à la droite de l'origine et d'autre part, entre tout ce qui se trouve à la gauche de l'origine, puis d'exprimer avec concision et précision le contraste entre les deux extrêmes.

L'interprétation des proximités entre les modalités devra aussi 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.

3.5.3.1 Etude des variables

Contributions des variables à l'inertie des axes

Nous savons que les contributions des variables à la formation de l'inertie du nuage dépendent essentiellement du nombre de leurs modalités. On peut cependant comparer leur contribution à l'inertie d'un axe à leur contribution à l'inertie du nuage, ce qui donne une idée de l'importance prise par chacune d'elles dans la formation des axes. Par exemple, nous obtenons ici :

Variable	Contribution à l'inertie du nuage	Contribution à l'inertie de l'axe 1	Contribution à l'inertie de l'axe 2	Contribution à l'inertie de l'axe 3
Mode d'occupation	0,1765	0,2420	0,2565	0,0754
Type de logement	0,2941	0,2142	0,3168	0,4467
Ancienneté	0,1765	0,2272	0,0988	0,0090
Eloignement	0,1176	0,1326	0,0049	0,0205
Superficie	0,2353	0,1840	0,3230	0,4484

On voit sur ce tableau que la part des variables "Mode d'occupation" et "Ancienneté" dans la formation du premier axe est supérieure à leur part dans l'inertie totale du nuage. De même pour les variables "Mode d'occupation" et "Type de logement" et "Superficie" pour le deuxième axe (alors que, pendant le même temps, la variable "Eloignement" ne joue pratiquement aucun rôle). Sur l'axe 3, les variables prédominantes sont "Type de logement" et "Superficie".

Constitution du tableau précédent :

Dans le tableau des résultats relatifs aux modalités, on additionne les inerties des différentes modalités d'une même variable, pour le nuage entier d'une part, et pour chacun des axes retenus d'autre part. Par exemple, pour la première variable :

	Inertie	Inertie dim 1	Inertie dim 2	Inertie dim 3
MODE:Seul	0,0304	0,0641	0,0661	0,0170
MODE:Coloc	0,0507	0,0020	0,0714	0,0521
MODE:Couple	0,0511	0,0018	0,1120	0,0048
MODE:Parents et NR	0,0442	0,1741	0,0070	0,0015
Total	**0,1764**	**0,2420**	**0,2565**	**0,0754**

Rapport de corrélation entre une variable factorielle et une question : questions structurantes pour un axe

Etant donné une question et un axe factoriel, le rapport de corrélation η^2 est défini par Escoffier et Pagès comme le quotient $\eta^2 = \frac{Inertie\,entre\,groupes}{Inertie\,totale}$, où l'inertie est calculée le long de l'axe considéré, et *les groupes sont définis par les modalités de la question considérée*. Ce coefficient mesure la liaison entre le facteur et la variable qualitative (question).
On peut montrer que ce rapport peut également être calculé par la formule :

$$\eta^2 = \text{Contribution de la variable à l'inertie de l'axe} \times \text{Valeur propre de l'axe} \times \text{Nombre de questions}$$

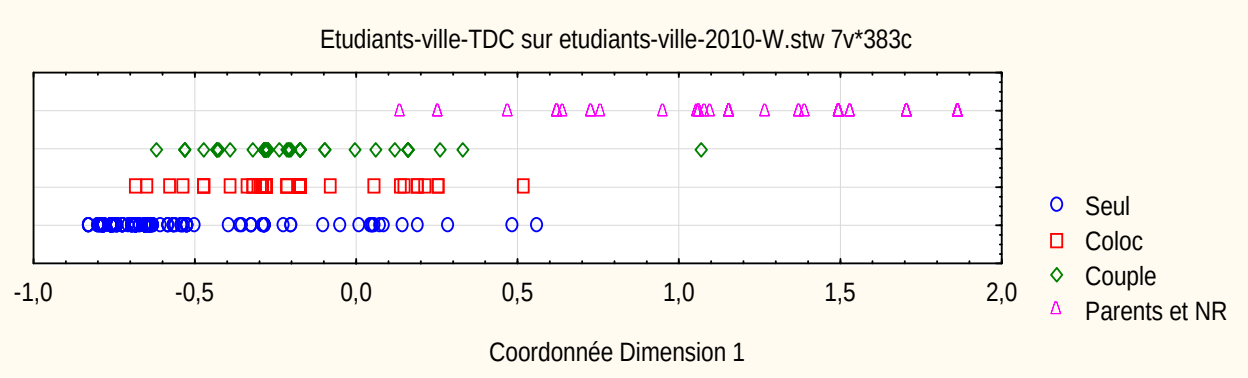
Ainsi, dans notre exemple :

$$\eta^2(\text{Axe}1, \text{Question}1) = 0{,}2420 \times 0{,}7215 \times 5 = 0{,}87$$
$$\eta^2(\text{Axe}2, \text{Question}1) = 0{,}2565 \times 0{,}4033 \times 5 = 0{,}52$$

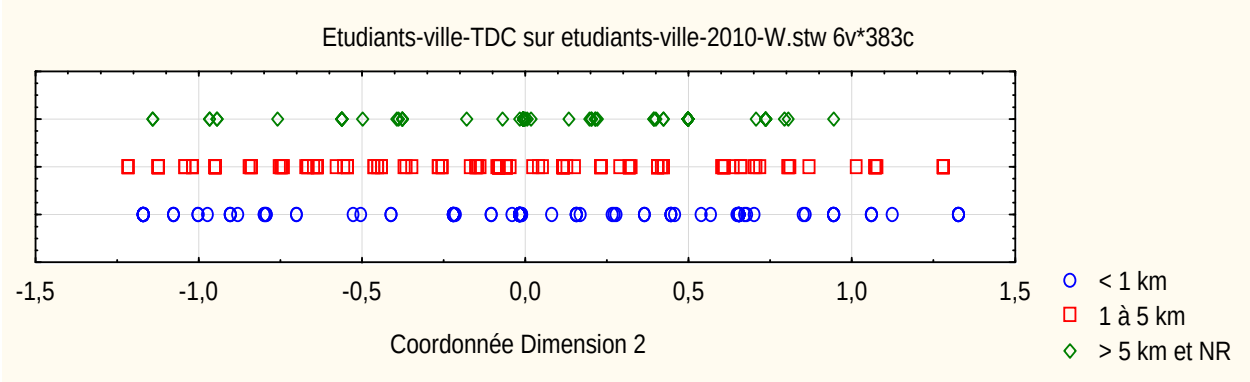
L'interprétation que l'on peut donner de ce rapport est la suivante : lorsque le rapport de corrélation est proche de 1, les individus d'une même classe sont très regroupés et les classes sont séparées les unes des autres ; c'est une situation de liaison très forte entre la variable qualitative et la variable numérique. Lorsqu'il est proche de 0, les moyennes des classes sont très proches de la moyenne générale et les individus d'une même classe sont très dispersés : la variable qualitative et la variable numérique ne sont pas liées. Sur notre exemple, les coefficients pour les 3 premiers axes sont :

	Axe 1	Axe 2	Axe 3
Mode d'occupation	0,87	0,52	0,12
Type de logement	0,77	0,64	0,73
Ancienneté	0,82	0,20	0,01
Eloignement	0,48	0,01	0,03
Superficie	0,66	0,65	0,74

Exemple : On a représenté les 383 sujets par leur coordonnée sur le premier axe, en repérant à l'aide de symboles différents les 4 modalités de la question "Mode d'occupation". Les quatre classes ainsi définies sont assez bien séparées sur le premier axe factoriel. En particulier, la modalité "Parents et NR" est bien distincte des autres modalités.

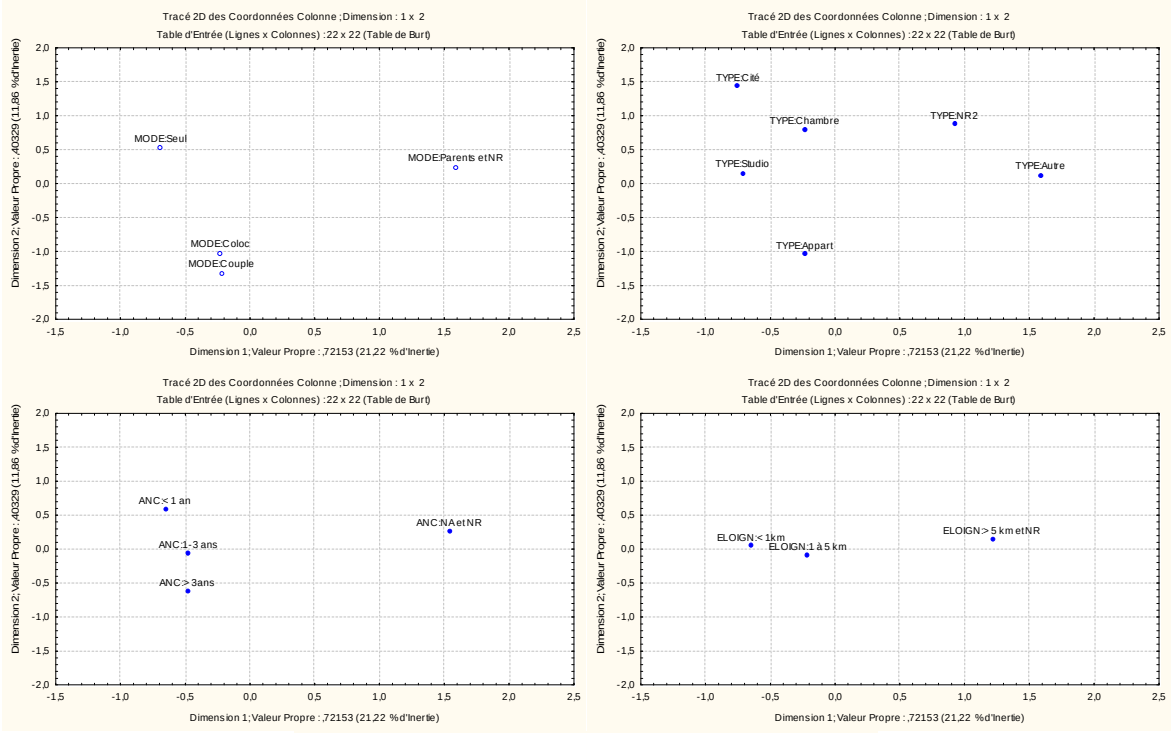


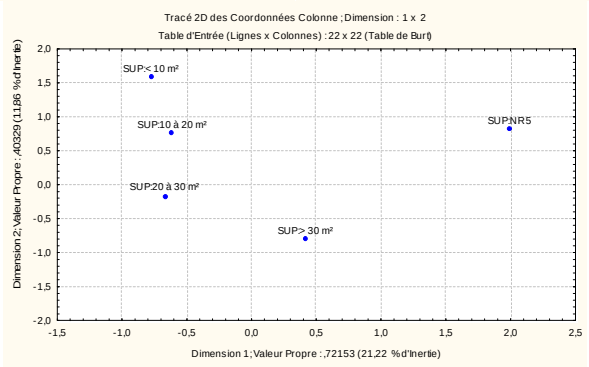
En revanche, une représentation analogue, faite sur l'axe 2 et en distinguant les 3 classes définies par les modalités de la question "Eloignement" montre l'absence de lien entre la variable qualitative "Eloignement" et la variable numérique "Coordonnée sur l'axe factoriel 2" :



Graphiques "par variable"

Dans certains cas, il peut être intéressant de réaliser des graphiques montrant la disposition des modalités d'une variable par rapport à 2 axes factoriels. Pour notre exemple, les modalités de chacune des variables se distribuent selon les deux premiers axes de la façon suivante :





3.5.3.2 Etude des axes

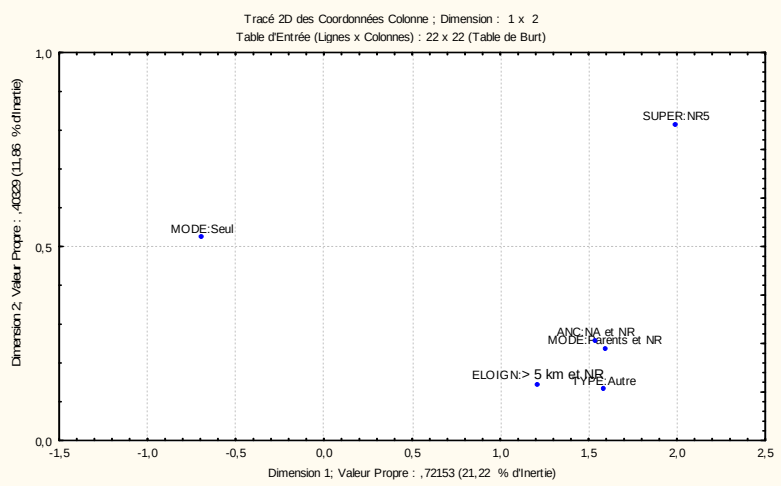
Pour chacun des axes, on pourra repérer les modalités dont la contribution à la formation de l'axe est supérieure à la moyenne (ici, 1/22=0,045), et éventuellement réaliser un graphique limité à ces seules modalités.

Premier Axe

Ainsi, pour le premier axe, on obtient :

-	+
MODE: Seul (6,41%)	MODE: Parents / NR (17,4%) ANC: NA et NR (16,8%) TYPE: Autre (13,8%) SUPER: NR5 (10,1%) ELOIGN:>5km (9,5%)

Représentation dans le premier plan factoriel des 6 modalités précédentes :

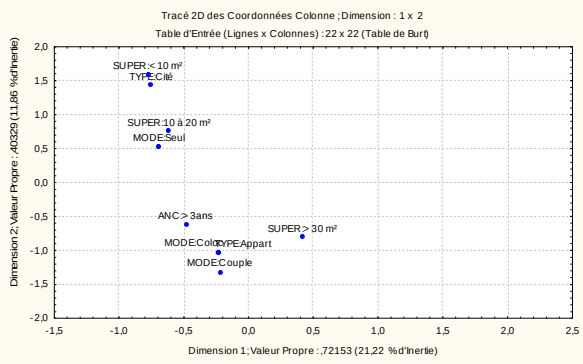


On voit que 4 des 5 variables figurent dans ce tableau, mais que les modalités concernées, dans la partie positive de l'axe sont essentiellement NA, NR, Autre et "Parents" pour la variable "MODE". A l'opposé, avec une contribution plus modeste, on trouve la modalité "Seul" de la variable "MODE". Cet axe semble opposer les étudiants logeant au foyer familial (les modalités telles que NA, Autre ou SUP:NR5 les concernent dans une large mesure) aux étudiants ayant un logement indépendant. Mais l'effet des modalités "non réponse" à faible effectif est sans doute aussi à prendre en compte.

Second axe

Pour le second axe, on obtient :

-	+
TYPE: Appart. (16,4%) SUPER: > 30 m^2 (12,4%) MODE: Couple (11,2%) MODE: Coloc. (7,1%) ANC: >3ans (5,3%)	SUPER: < 10m^2 (11,5%) TYPE: Cité (11%) MODE: Seul (6,6%) SUPER: 10 à 20 m^2 (4,99%)

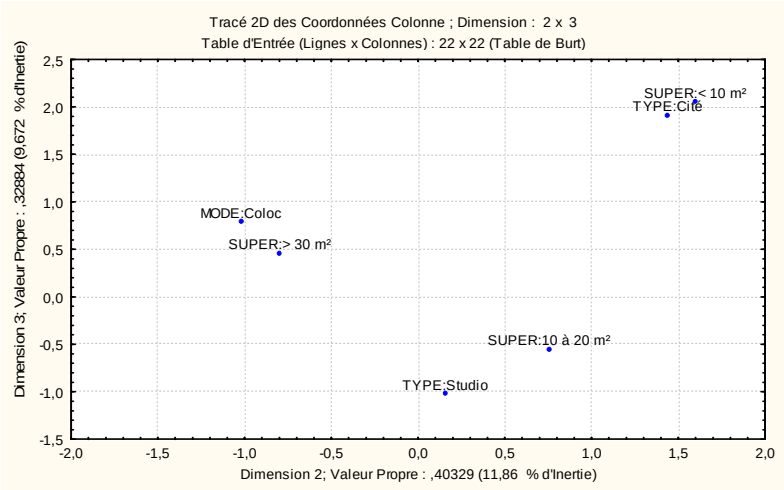


Cet axe fait essentiellement intervenir les variables TYPE, SUPERFICIE et MODE. Les modalités mises en jeu concernent essentiellement les étudiants qui n'habitent plus au foyer familial. L'axe oppose clairement les étudiants vivant seuls en cité universitaire, dans un logement de faible superficie (partie positive de l'axe) aux étudiants vivant en couple ou en colocation, en appartement, de superficie plus importante (partie négative de l'axe).

Troisième axe

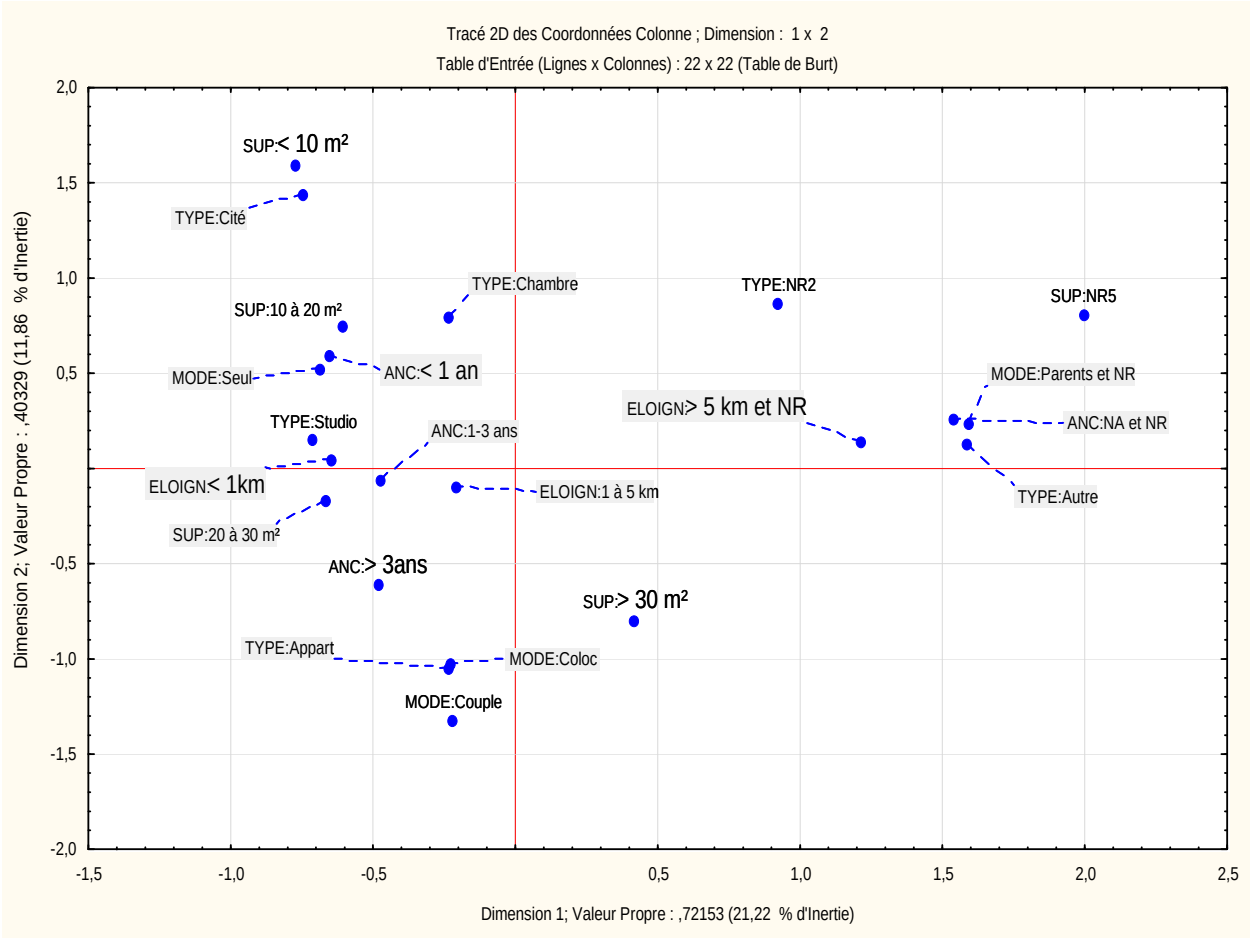
Pour le 3ème axe, on obtient :

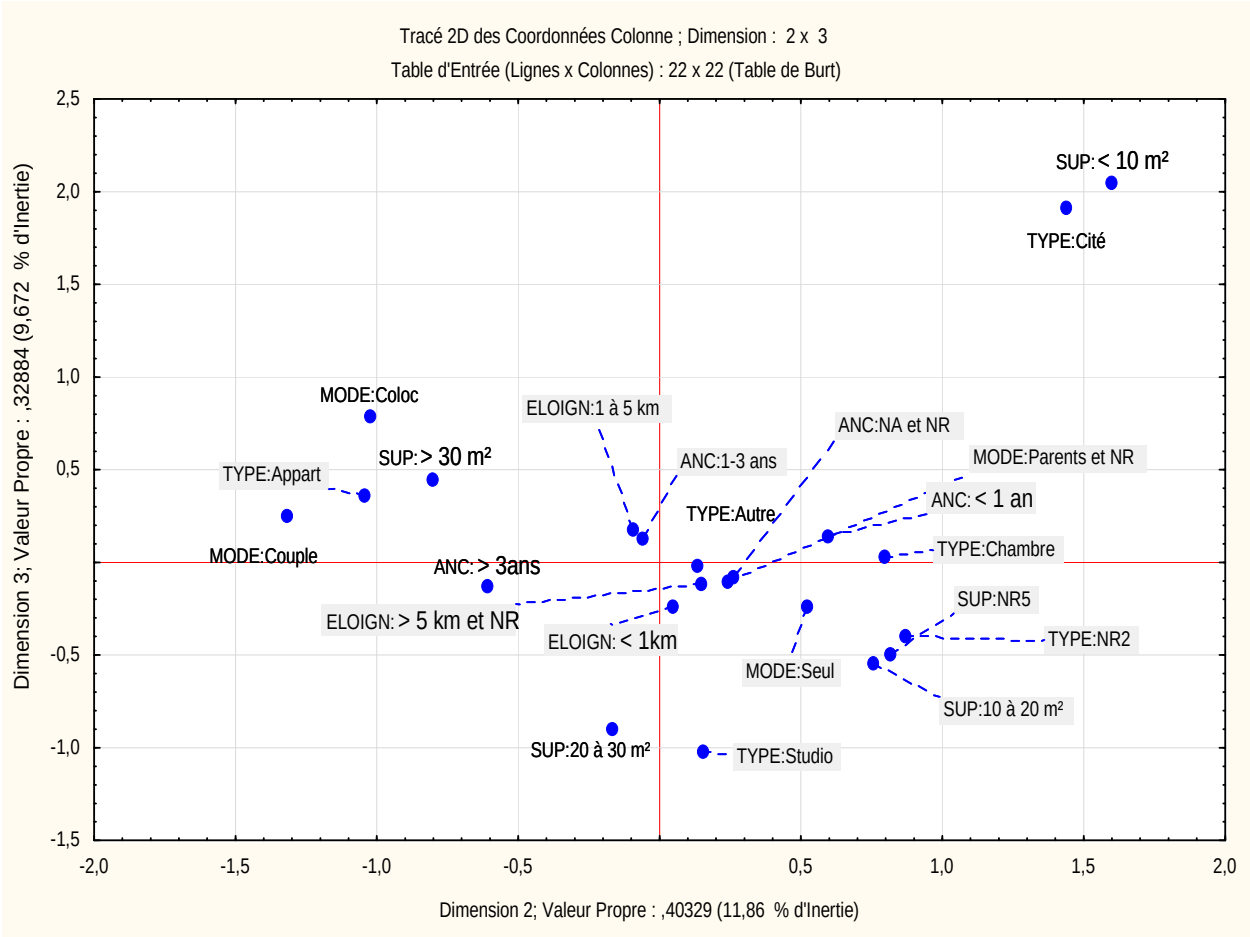
-	+
TYPE: Studio (18,0%) SUPER: 20 à 30m^2 (12,1%)	TYPE: Cité (23,8%) SUPER: <10m^2 (23,3%) MODE: Coloc (5,2%) SUPER: >30m^2 (4,85%)



C'est essentiellement la superficie, et le type de logement correspondant, qui interviennent ici : logement en cité universitaire, de moins de 10 m^2, studio de 20 à 30 m^2 et colocation (plutôt en appartement) de plus de 30 m^2. Encore une fois, les étudiants habitant au domicile familial interviennent peu dans la formation de cet axe.

3.5.3.3 Graphiques selon les deux premiers plans principaux

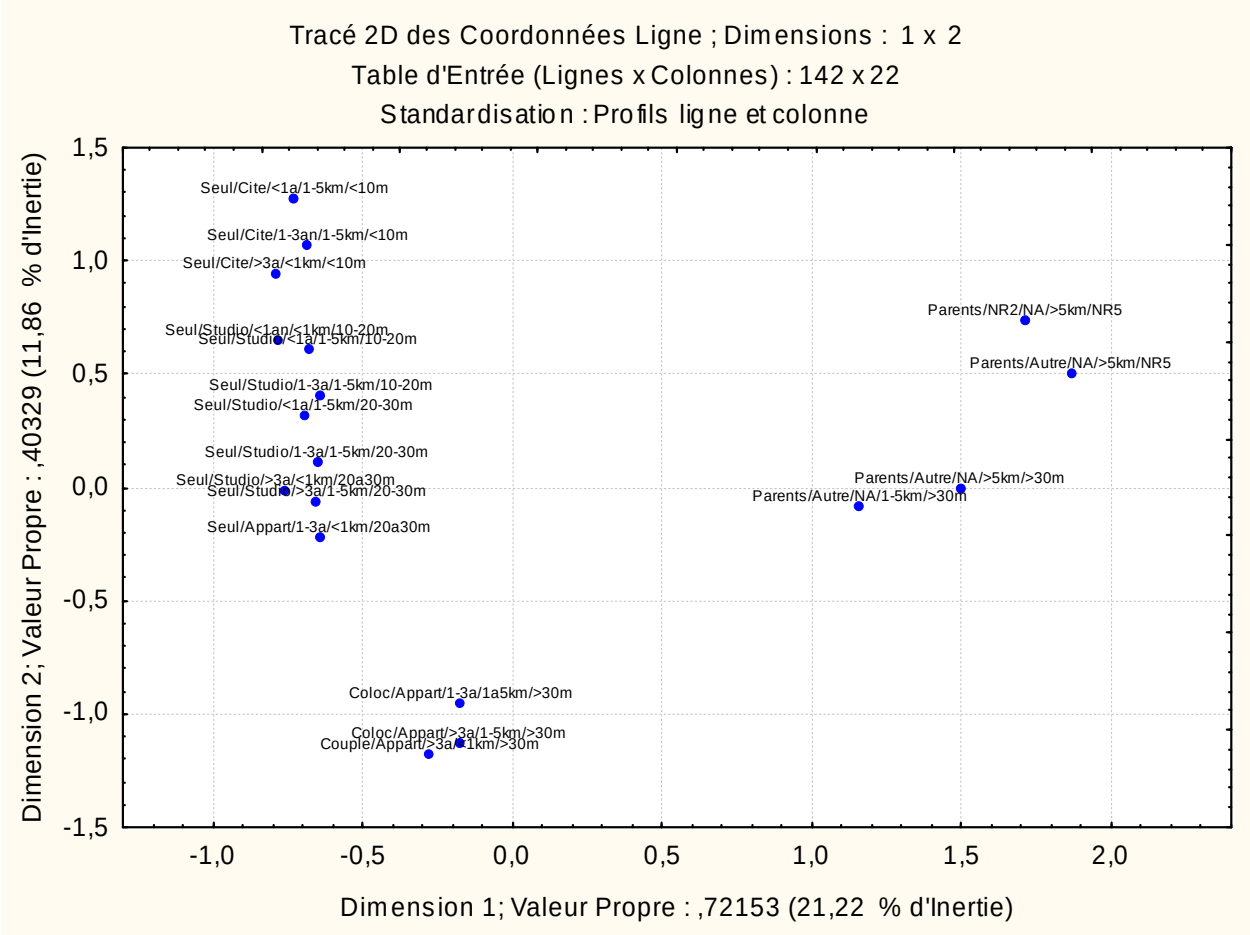




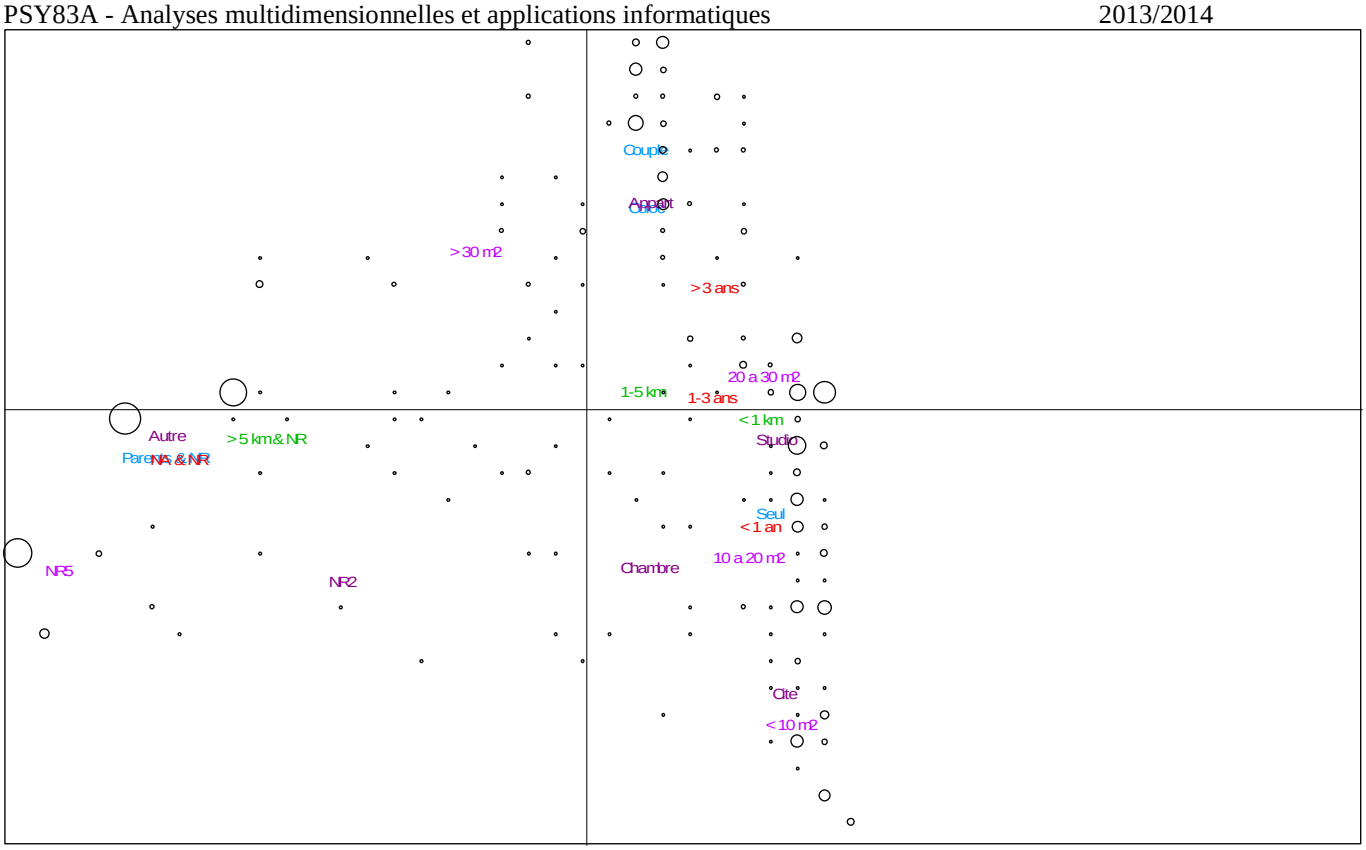
L'étude précédente a permis de distinguer essentiellement trois groupes : le groupe des étudiants logeant chez leurs parents, et des non réponses (partie positive du premier axe), le groupe des étudiants vivant seuls en cité ou en chambre (partie positive du deuxième axe) et le groupe des étudiants vivant en couple ou en colocation (partie négative du deuxième axe. Les étudiants vivant en studio constituent un groupe intermédiaire entre les deux précédents. On voit que les variables "Eloignement" et "Ancienneté" ont joué des rôles assez modestes dans l'étude. L'ancienneté de moins d'un an, et un éloignement faible sont plutôt associés à l'habitat en cité universitaire alors que l'habitat de type "studio" ou "colocation" est associé à une ancienneté et une distance plus importantes. Enfin, les distances supérieures à 5 km se rencontreraient plutôt chez les étudiants logeant chez leurs parents.

3.5.3.4 Etude des patrons de réponses

L'étude des patrons de réponses présente ici un intérêt limité, puisque le protocole a été généré artificiellement à partir du tableau de Burt. Le positionnement dans le premier plan factoriel des 18 patrons les plus fréquents confirme cependant ce qui a été obtenu à partir des modalités :



On pourra comparer ce graphique à celui produit par un autre logiciel (Modalisa) dans lequel les patrons de réponses sont représentés par des cercles proportionnels à l'effectif :



3.6 L'ACM avec Statistica

3.6.1 Procédure à utiliser selon la forme des données d'entrée

On a observé plusieurs (2, 3 ou plus) variables nominales sur une population, et on souhaite explorer ces données à l'aide d'une ACM. Mais on sait que l'ACM est en fait une AFC particulière (AFC du tableau disjonctif complet). Nous allons voir que, selon la forme sous laquelle ces données sont disponibles, on utilise sous Statistica les menus suivants :

Format des données	Onglet "Analyse des Correspondances"	Onglet "Analyse des Correspondances Multiple"	Observations
Tableau protocole	Non, si plus de 2 variables	Oui	AFC impossible si plus de 2 variables
Tableau d'effectifs	Non, si plus de 2 variables	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 fournissent des résultats analogues, mais pas identiques

3.6.2 Présentation des données étudiées

Référence : Les données présentées ici sont accessibles sur le site personnel de Gilles Hunault, à l'adresse :
http://www.info.univ-angers.fr/pub/gh/Datasets/pbio.htm.

Ce dossier contient des données relatives à une enquête réalisée dans des supermarchés angevins et parisiens entre 1996 et 1998 dans le but de connaître l'avis de consommateurs quant aux produits biologiques et aux produits diététiques.

La structure des données est la suivante :

1 - Matricule anonyme de la personne interrogée
2 - Connaissez-vous les produits biologiques ?
- 0 - non réponse
- 1 - oui
- 2 - non

3 - Y a-t-il une différence entre produit biologique et produit diététique ?
- 0 - non réponse
- 1 - oui
- 2 - non

4 - Avez-vous déjà consommé des produits biologiques ?
- 1 - non jamais
- 2 - oui une seule fois
- 3 - oui rarement
- 4 - oui de temps en temps
- 5 - oui plusieurs fois par mois
- 6 - oui plusieurs fois par semaine
- 7 - ne se prononce pas

5 - Parmi les marques suivantes lesquelles connaissez-vous ?
- 0 - non réponse
- 1 - bio vivre
- 2 - bjorg
- 3 - carrefour bio
- 4 - la vie
- 5 - vrai
- 6 - prosain
- 7 - favrichon

6 - Avez-vous déjà consommé des produits " La Vie " ?
- 0 - non réponse
- 1 - oui une fois
- 2 - oui occasionnellement
- 3 - oui régulièrement
- 4 - non jamais

7 - Sexe de la personne
- 1 - homme
- 2 - femme

8 - Classe d'age
- 1 - moins de 25 ans
- 2 - entre 25 et 35 ans
- 3 - entre 35 et 45 ans
- 4 - entre 45 et 55 ans
- 5 - entre 55 et 65 ans
- 6 - plus de 65 ans

9 -Etat-civil
- 0 - autre
- 1 - marie
- 2 - célibataire
- 3 - divorcé
- 4 - en concubinage
- 5 - veuf

10 - Nombre d'enfants
- 1 - 0 enfant
- 2 - 1 enfants
- 3 - 2 enfants
- 4 - 3 enfants
- 5 - plus de 3 enfants

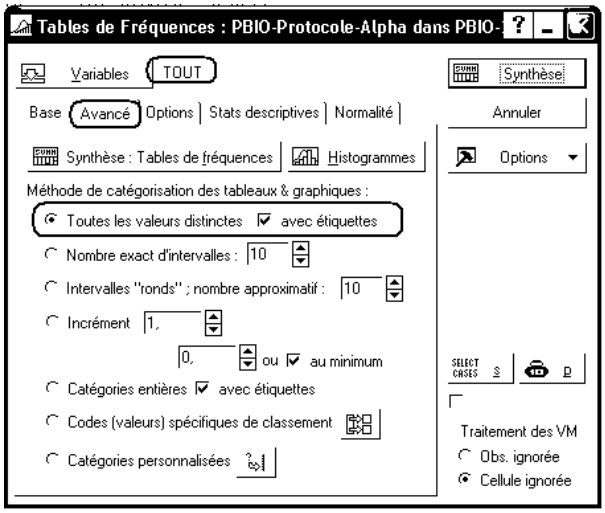
11 - Situation professionnelle
- 0 - non-réponse
- 1 - agriculteur
- 2 - artisan
- 3 - cadre supérieur
- 4 - cadre moyen
- 5 - employé
- 6 - ouvrier
- 7 - retraité
- 8 - autre

12 - Classe de revenus mensuels
- 0 - non réponse
- 1 - moins de 5 kF
- 2 - entre 5 et 10 kF
- 3 - entre 10 et 15 kF
- 4 - entre 15 et 20 kF
- 5 - plus de 20 kF
- 6 - ne se prononce pas

L'échantillon interrogé comporte 419 observations. Les données figurent (sous plusieurs formes) dans le classeur PBIO-2010.stw du serveur de TD. En particulier, le protocole observé se trouve dans la feuille PBIO-Protocole-Num (les modalités y sont représentées par leur code numérique) et dans la feuille PBIO-Protocole-Alpha (dans cette feuille, des étiquettes de texte, représentant des abréviations des libellés de réponses, ont été introduites).

3.6.3 Exploration préalable des données

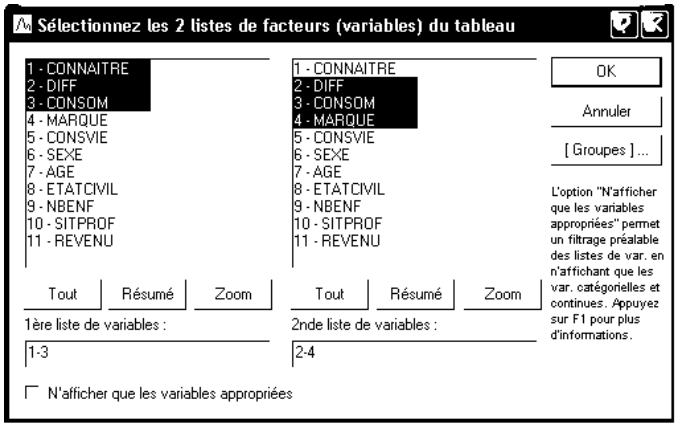
Avant de réaliser des analyses multivariées sur ces données, explorez-les en utilisant le menu Statistiques - Statistiques Elémentaires - Tables de fréquences pour obtenir les tris à plat des différentes variables. On obtiendra l'ensemble des tris à plat des différentes variables en complétant le dialogue de la façon suivante :



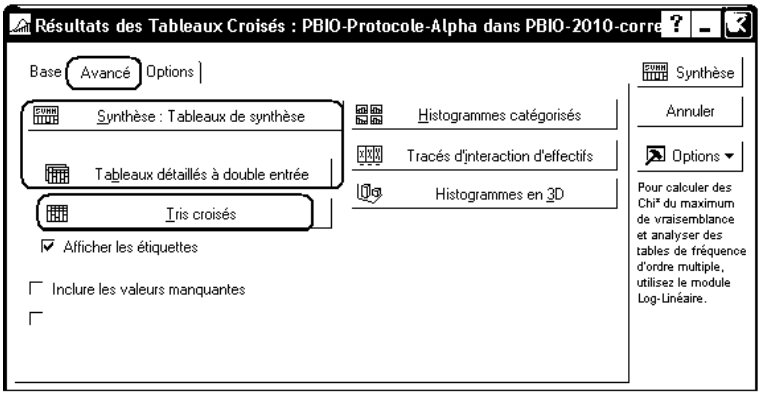
Par exemple, pour la variable DIFF, on obtient :

Table de fréquences : DIFF (PBIO-2.sta)				
	Effectif	Effectifs Cumulés	%age	%age Cumulé
NR	3	3	0,71599	0,7160
oui	328	331	78,28162	78,9976
non	88	419	21,00239	100,0000
VM	0	419	0,00000	100,0000

On peut également utiliser le menu Statistiques - Statistiques Elémentaires - Statistiques Descriptives - Tableaux et tris croisés pour obtenir des tableaux de contingence croisant les variables deux à deux. Par exemple, on peut utiliser l'onglet Tris Croisés. On indique deux listes de variables, une même variable pouvant être présente dans chacune des deux listes. Par exemple :



Après avoir cliqué sur le bouton "OK", on a accès au dialogue des résultats. Dans l'onglet "Avancé", le bouton "Synthèse" et le bouton "Tableaux détaillés à double entrée" génèrent autant de feuilles de résultats que de couples de variables obtenus en croisant les deux listes. Le bouton "Tri croisés" permet d'obtenir les résultats rassemblés dans un seul tableau, analogue à un tableau de Burt.



Exemple :

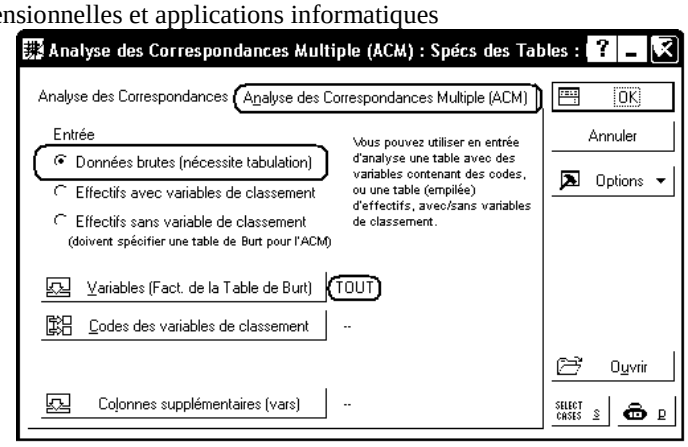
Table de Fréquences - Synthèse (PBIO-2.sta)								
Table :CONSOM(7) x AGE(6)								
	CONSOM	AGE	AGE	AGE	AGE	AGE	AGE	Totaux
		- de 25 ans	entre 25 et 35 ans	entre 35 et 45 ans	entre 45 et 55 ans	entre 55 et 65 ans	+ de 65 ans	Ligne
	jamais	15	39	19	21	9	8	111
	une fois	3	5	0	3	3	1	15
	rarement	16	25	17	19	4	8	89
	quelquefois	13	28	16	38	12	10	117
	souvent	2	4	3	10	2	3	24
	très souvent	0	11	13	8	9	8	49
	nr	3	4	2	2	2	1	14
	Ts Grpes	52	116	70	101	41	39	419

3.6.4 ACM menée à partir d'un tableau protocole

3.6.4.1 ACM à partir d'un tableau protocole : étude sur l'ensemble des variables

On peut essayer de réaliser une ACM avec, comme données actives, l'ensemble des données fournies. Gardez ici PBIO-Protocole-Alpha comme feuille active. Utilisez le menu Statistiques - Techniques exploratoires multivariées - Analyse des correspondances et sélectionnez ensuite l'item "Analyse des Correspondances Multiples".

On voit que l'analyse peut être menée soit à partir d'un tableau protocole, soit à partir d'un tableau d'effectifs, soit à partir d'un tableau de Burt. Dans notre cas, les données se présentent sous forme d'un tableau protocole.



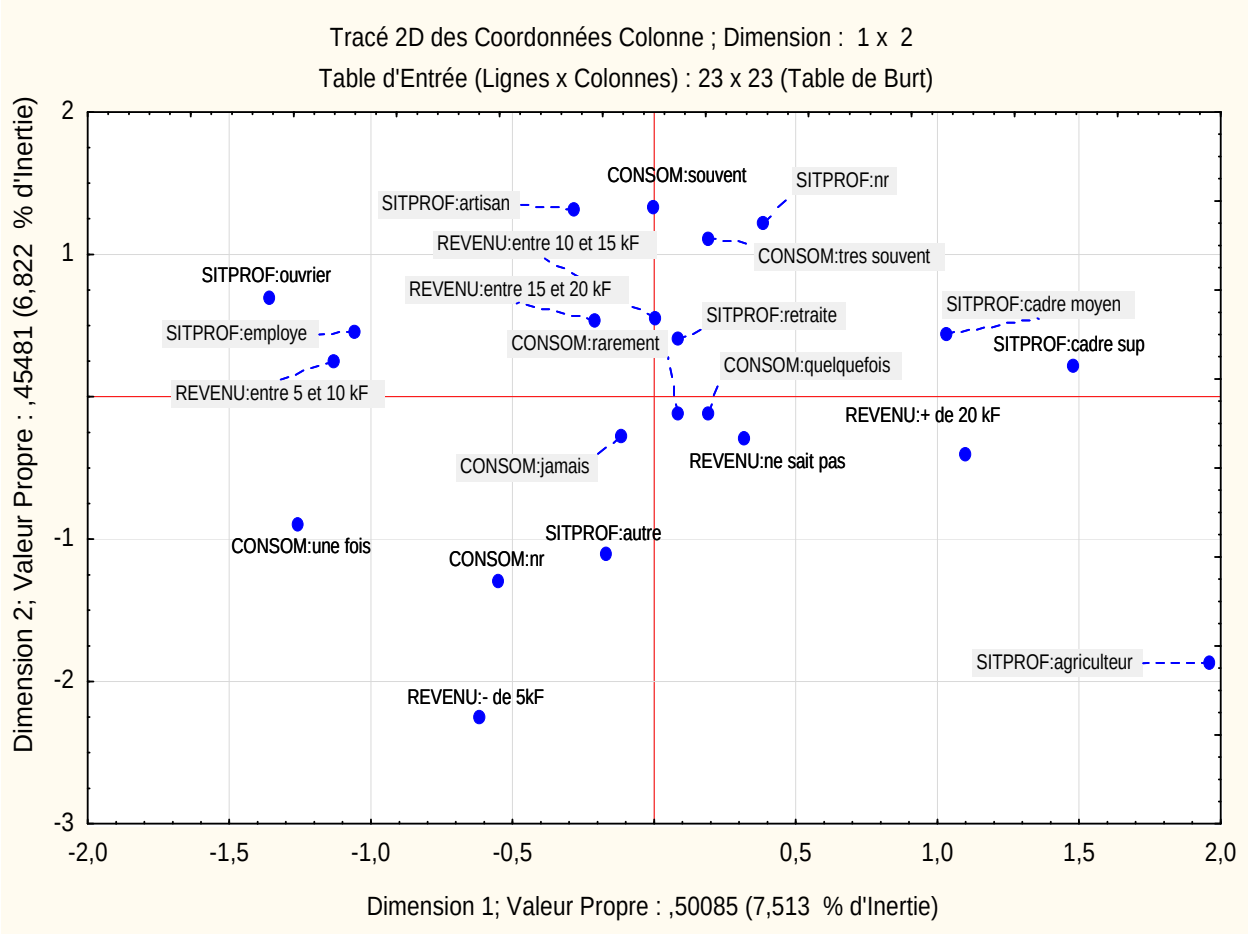
Cependant, les modalités sont alors trop nombreuses et les résultats sont difficilement interprétables. Par ailleurs, certains problèmes cités plus haut se rencontrent sur ces données :

- Certaines modalités ont des effectifs très faibles, et leur importance dans l'étude est surestimée;
- Le nombre de modalités de chaque variable varie de 3 à 9. Or, on sait qu'il vaut mieux utiliser des variables comportant à peu près le même nombre de modalités.

3.6.4.2 ACM menée sur les variables CONSOM, SITPROF et REVENU

Il est ici préférable de choisir 3 ou 4 variables, pour lesquelles nous pensons que l'étude conjointe présente un intérêt. Ce choix s'imposerait de lui-même si nous avions fixé un but bien déterminé à notre étude, ce qui n'est pas le cas ici. L'examen des tables de fréquences nous permet cependant de choisir des variables où les effectifs sont assez bien répartis entre les modalités, telles que, par exemple : CONSOM, SITPROF et REVENU.

En recadrant les échelles des axes (bouton droit sur l'axe, puis le menu local "Echelle...") de façon à éliminer certaines parties sans intérêt du dessin, on obtient le graphique suivant :



Pour l'essentiel, les résultats sont disposés de la même manière que ceux de l'AFC, mais ne concernent que les modalités colonnes. En effet, Statistica travaille à partir du tableau de Burt, qui est d'ailleurs accessible à l'aide du bouton "Effectifs Observés" de l'onglet "Etude". On peut également noter que les tableaux de pourcentages lignes et colonnes sont obtenus à partir du tableau de Burt.

Calcul des taux d'inertie modifiés

On sait que Statistica ne fournit pas directement ce résultat. Nous allons donc devoir recourir à Excel pour réaliser le calcul.
Dans Statistica, affichez le tableau des 20 valeurs propres correspondant à cette analyse. Sélectionnez l'ensemble du tableau et utilisez le menu Edition - Copier avec noms.
Sans quitter Statistica, ouvrez un classeur Excel et collez le contenu du tableau, à partir de la cellule A1.
Calculez ensuite successivement :

- la moyenne des valeurs propres : entrez par exemple la formule =MOYENNE(C2:C21) en C23
- pour les valeurs propres supérieures à cette moyenne, les valeurs propres modifiées (écart à la moyenne précédente élevé au carré) : entrez par exemple la formule =(C2-C$23)^2 en G2 et recopiez-la jusqu'en G11
- la somme de ces valeurs propres : entrez par exemple la formule =SOMME(G2:G11) en G23
- le taux d'inertie modifié correspondant : entrez par exemple la formule =G2/G$23 en H2 et recopiez-la jusqu'en H11.

Vous devriez aboutir à un tableau du type suivant :

	ValSing.	ValProp.	%age	%age	Chi²	VP modif.	Tx modifiés
1	0,7077	0,5008	7,5127	7,51	665,74	0,02806	42,04%
2	0,6744	0,4548	6,8222	14,33	604,55	0,01476	22,11%
3	0,6571	0,4317	6,4760	20,81	573,87	0,00968	14,51%
4	0,6435	0,4141	6,2119	27,02	550,47	0,00653	9,78%
5	0,6318	0,3992	5,9884	33,01	530,66	0,00434	6,50%
6	0,6171	0,3808	5,7116	38,72	506,13	0,00225	3,37%

7	0,6004	0,3604	5,4065	44,13	479,10	0,00073	1,10%
8	0,5938	0,3526	5,2896	49,42	468,74	0,00037	0,56%
9	0,5816	0,3382	5,0737	54,49	449,61	0,00002	0,04%
10	0,5779	0,3340	5,0103	59,50	443,99	0,00000	0,00%
11	0,5662	0,3206	4,8088	64,31	426,13		
12	0,5583	0,3117	4,6755	68,99	414,32		
13	0,5506	0,3031	4,5470	73,53	402,93		
14	0,5481	0,3004	4,5055	78,04	399,26		
15	0,5384	0,2899	4,3484	82,39	385,34		
16	0,5242	0,2747	4,1212	86,51	365,20		
17	0,5114	0,2616	3,9235	90,43	347,68		
18	0,4919	0,2419	3,6288	94,06	321,56		
19	0,4853	0,2355	3,5321	97,59	313,00		
20	0,4005	0,1604	2,4062	100,00	213,22		
	Moyenne	0,3333			Total	0,06675	

Exploitation de ce résultat : après transformation, il reste 10 valeurs propres. Parmi celles-ci, les 3 premières (et éventuellement la 4ème) ont des taux d'inertie modifiés supérieurs ou voisins de 100%/10 = 10%. L'étude devrait donc se concentrer sur les 3, voire 4 premiers axes.

Tableau des résultats relatifs aux lignes

L'interprétation utilisera également les tableaux des coordonnées, qualités de représentation et contributions :

NomLigne	Coordonnées Colonne et Contributions à l'Inertie (PBIO-2.sta)									
	Inertie Totale = 6,6667									
	Ligne Numéro	Coord. Dim.1	Coord. Dim.2	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus2 Dim.1	Inertie Dim.2	Cosinus2 Dim.2
CONSOM:jamais	1	-0,1147	-0,2745	0,0883	0,0319	0,0368	0,0023	0,0047	0,0146	0,0271
CONSOM:une fois	2	-1,2597	-0,8938	0,0119	0,0886	0,0482	0,0378	0,0589	0,0210	0,0297
CONSOM:rarement	3	0,0862	-0,1216	0,0708	0,0060	0,0394	0,0011	0,0020	0,0023	0,0040
CONSOM:quelquefois	4	0,1921	-0,1167	0,0931	0,0196	0,0360	0,0069	0,0143	0,0028	0,0053
CONSOM:souvent	5	-0,0048	1,3379	0,0191	0,1088	0,0471	0,0000	0,0000	0,0751	0,1088
CONSOM:très souvent	6	0,1899	1,1087	0,0390	0,1676	0,0442	0,0028	0,0048	0,1054	0,1628
CONSOM:nr	7	-0,5510	-1,2917	0,0111	0,0682	0,0483	0,0068	0,0105	0,0409	0,0577
SITPROF:nr	8	0,3839	1,2134	0,0008	0,0039	0,0499	0,0002	0,0004	0,0026	0,0035
SITPROF:agriculteur	9	1,9600	-1,8723	0,0008	0,0176	0,0499	0,0061	0,0092	0,0061	0,0084
SITPROF:artisan	10	-0,2851	1,3171	0,0072	0,0399	0,0489	0,0012	0,0018	0,0273	0,0381
SITPROF:cadre sup	11	1,4783	0,2187	0,0270	0,1972	0,0459	0,1180	0,1930	0,0028	0,0042
SITPROF:cadre moyen	12	1,0310	0,4370	0,0676	0,3191	0,0399	0,1435	0,2705	0,0284	0,0486
SITPROF:employé	13	-1,0596	0,4535	0,0835	0,4442	0,0375	0,1873	0,3755	0,0378	0,0688
SITPROF:ouvrier	14	-1,3569	0,7003	0,0064	0,0454	0,0490	0,0234	0,0358	0,0069	0,0095
SITPROF:retraité	15	0,0832	0,4072	0,0453	0,0272	0,0432	0,0006	0,0011	0,0165	0,0261
SITPROF:autre	16	-0,1707	-1,1109	0,0947	0,5011	0,0358	0,0055	0,0116	0,2569	0,4895
REVENU:nr	17	2,9320	4,2712	0,0008	0,0642	0,0499	0,0137	0,0206	0,0319	0,0436
REVENU:- de 5kF	18	-0,6142	-2,2576	0,0175	0,3034	0,0474	0,0132	0,0209	0,1961	0,2824
REVENU:entre 5 et 10 kF	19	-1,1327	0,2443	0,0788	0,4154	0,0382	0,2018	0,3969	0,0103	0,0185
REVENU:entre 10 et 15 kF	20	0,0009	0,5590	0,0628	0,0726	0,0406	0,0000	0,0000	0,0432	0,0726
REVENU:entre 15 et 20 kF	21	-0,2119	0,5275	0,0525	0,0604	0,0421	0,0047	0,0084	0,0321	0,0520
REVENU:+ de 20 kF	22	1,0996	-0,4079	0,0899	0,5080	0,0365	0,2170	0,4465	0,0329	0,0614
REVENU:ne sait pas	23	0,3174	-0,2993	0,0310	0,0195	0,0453	0,0062	0,0103	0,0061	0,0092

En recopiant ce tableau dans Excel, on peut également calculer les inerties relatives des questions et leurs contributions à l'inertie de chacun des axes, ainsi que les rapports η^2 définis au paragraphe 3.5.3.1.

Rappel : On montre que ce rapport peut être calculé par la formule :

$\eta^2 = \text{Contribution de la variable à l'inertie de l'axe} \times \text{Valeur propre de l'axe} \times \text{Nombre de questions}$

	Inr des	Inr Dim 1	Eta-2 Axe 1	Inr Dim 2	Eta-2 Axe 2

	variables				
CONSO	0,30	0,0576	0,0865	0,2620	0,3575
SITPROF	0,40	0,4858	0,7300	0,3853	0,5257
REVENU	0,30	0,4566	0,6860	0,3527	0,4812
	Valeur propre	0,500846		0,454813	

Après avoir collé le tableau précédent à partir de la cellule A1, dans la plage A1:K24, on pourra par exemple utiliser les formules suivantes :

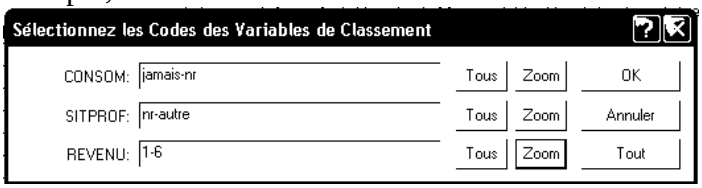
	F	G	H	I	J	K
25		Inr des variables	Inr Dim 1	Eta-2 Axe 1	Inr Dim 2	Eta-2 Axe 2
26	CONSO	=SOMME(G2:G8)	=SOMME(H2:H8)	=H26*H$30*3	=SOMME(J2:J8)	=J26*J$30*3
27	SITPROF	=SOMME(G9:G17)	=SOMME(H9:H17)	=H27*H$30*3	=SOMME(J9:J17)	=J27*J$30*3
28	REVENU	=SOMME(G18:G24)	=SOMME(H18:H24)	=H28*H$30*3	=SOMME(J18:J24)	=J28*J$30*3
29						
30		Valeur propre	0,500846		0,454813	

On voit par exemple qu'il existe un coefficient de corrélation important entre les coordonnées des individus le long du premier axe et la question SITPROF. Autrement dit, les sujets correspondant à une même modalité de la question SITPROF sont bien regroupés et les différents groupes définis par ces modalités sont bien séparés les uns des autres.

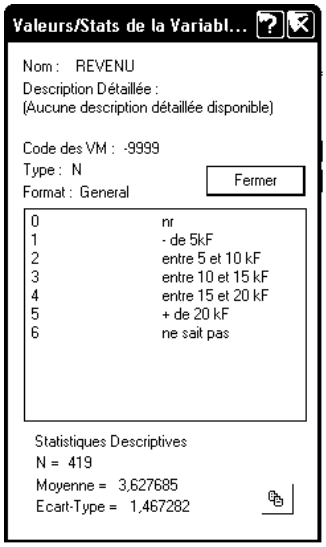
3.6.4.3 Elimination des modalités de faible effectif

L'étude précédente montre que les modalités "non réponse", notamment la modalité "nr" de la variable "REVENU", occupent des positions excentrées sur le graphique. Lorsqu'on travaille à partir d'un tableau protocole, on peut facilement éliminer ces modalités de la façon suivante :

- On sélectionne les variables comme précédemment
- On utilise ensuite le bouton "Codes des variables de classement" et l'on complète le dialogue, par exemple, comme suit :



On pourra utiliser le bouton "Zoom" pour déterminer les codes numériques des modalités à sélectionner :



On pourra observer que le fait d'éliminer cette modalité ne modifie pas vraiment les résultats produits, mais améliore la lisibilité des graphiques.

3.6.5 ACM menée à partir d'un tableau de Burt

Statistica permet également de générer un tableau de Burt à partir du protocole et/ou de réaliser une ACM à partir d'un tableau de Burt. On peut pour cela utiliser la feuille de données PBIO-Burt-CONSO-AGE-ETATCIVIL du classeur ou générer un tableau de Burt selon la méthode indiquée dans le paragraphe ci-dessous.

3.6.5.1 Générer le tableau de Burt correspondant aux variables CONSOM, AGE et ETAT CIVIL

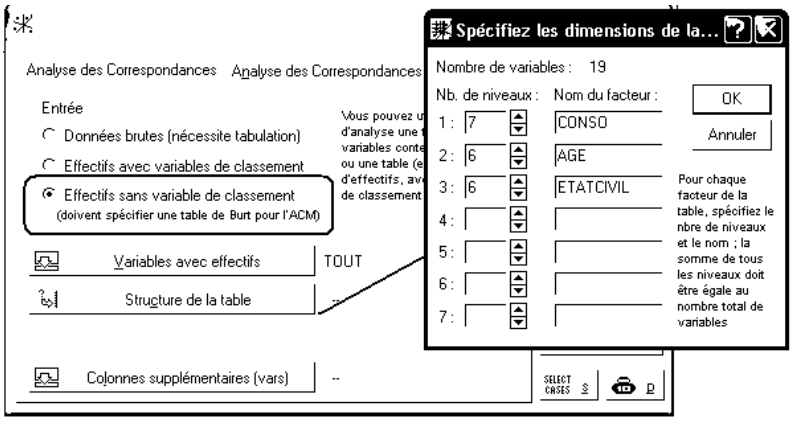
Faites une nouvelle Analyse des correspondances multiples en sélectionnant les variables CONSOM, AGE et ETAT CIVIL comme variables de l'étude. Sélectionnez ensuite l'onglet "Etude" puis le bouton "Effectifs observés".

La feuille de résultats obtenue est constituée du tableau de Burt relatif à ces trois variables, et des marges ligne et colonne correspondantes. On peut ensuite la copier (menu local "Copier le document du classeur") et la coller dans le classeur, au même niveau hiérarchique que les autres feuilles de données. On peut également supprimer la dernière ligne, la dernière colonne et renommer les variables, faute de quoi, toutes les modalités d'une même question seront étiquetées de façon identique dans le graphique produit. On obtient ainsi un résultat tel que :

	jamais	une fois	rarement	quelquefois	souvent	très souvent	NR	- de 25 ans	de 25 à 35 ans	de 35 à 45 ans	de 45 à 55 ans	de 55 à 65 ans	+ de 65 ans	autre	marié	célibataire	divorcé	concubinage	veuf
CONSOM:jamais	111	0	0	0	0	0	0	15	39	19	21	9	8	3	58	30	4	12	4
CONSOM:une fois	0	15	0	0	0	0	0	3	5	0	3	3	1	0	8	5	0	1	1
CONSOM:rarement	0	0	89	0	0	0	0	16	25	17	19	4	8	0	45	32	3	7	2
CONSOM:quelquefois	0	0	0	117	0	0	0	13	28	16	38	12	10	1	72	25	9	7	3
CONSOM:souvent	0	0	0	0	24	0	0	2	4	3	10	2	3	1	9	7	2	4	1
CONSOM:tres souvent	0	0	0	0	0	49	0	0	11	13	8	9	8	0	30	9	2	6	2
CONSOM:nr	0	0	0	0	0	0	14	3	4	2	2	2	1	1	6	6	1	0	0
AGE:- de 25 ans	15	3	16	13	2	0	3	52	0	0	0	0	0	1	4	41	0	6	0
AGE:entre 25 et 35 ans	39	5	25	28	4	11	4	0	116	0	0	0	0	3	45	46	2	19	1
AGE:entre 35 et 45 ans	19	0	17	16	3	13	2	0	0	70	0	0	0	0	41	15	4	8	2
AGE:entre 45 et 55 ans	21	3	19	38	10	8	2	0	0	0	101	0	0	2	81	5	6	4	3
AGE:entre 55 et 65 ans	9	3	4	12	2	9	2	0	0	0	0	41	0	0	31	4	5	0	1
AGE:+ de 65 ans	8	1	8	10	3	8	1	0	0	0	0	0	39	0	26	3	4	0	6
ETATCIVIL: autre	3	0	0	1	1	0	1	1	3	0	2	0	0	6	0	0	0	0	0
ETATCIVIL: marie	58	8	45	72	9	30	6	4	45	41	81	31	26	0	228	0	0	0	0
ETATCIVIL: celibataire	30	5	32	25	7	9	6	41	46	15	5	4	3	0	0	114	0	0	0
ETATCIVIL: divorce	4	0	3	9	2	2	1	0	2	4	6	5	4	0	0	0	21	0	0
ETATCIVIL: concubinage	12	1	7	7	4	6	0	6	19	8	4	0	0	0	0	0	0	37	0
ETATCIVIL: veuf	4	1	2	3	1	2	0	0	1	2	3	1	6	0	0	0	0	0	13

3.6.5.2 ACM menée variables CONSOM, AGE et ETAT CIVIL

La différence la plus importante par rapport à la situation précédente est la nécessité d'indiquer à Statistica les variables qui doivent être regroupées afin de former un "facteur" de l'étude. Ici, "CONSOM" correspond aux 7 premières variables, "AGE" aux 6 suivantes et "ETAT CIVIL" aux 6 dernières :



Il faut noter que Statistica exige que le tableau fourni soit rigoureusement un tableau de Burt. Il n'est pas possible, par exemple, de supprimer les lignes correspondant aux non-réponses, car la cohérence du tableau n'est alors plus assurée.

3.6.5.3 Colonnes supplémentaires dans un tableau de Burt

Pour l'ACM comme pour l'ACP et l'AFC, il est possible d'indiquer certains éléments comme éléments supplémentaires, afin qu'ils figurent dans les résultats sans avoir influencé le calcul proprement dit. Lorsque l'ACM est faite à partir d'un tableau de Burt :

- Il faut indiquer toutes les variables comme "Variables avec Effectifs" et préciser la structure de la table, aussi bien pour les colonnes actives que pour les colonnes supplémentaires ;
- Les colonnes sélectionnées sous la rubrique "Colonnes supplémentaires" doivent correspondre à la totalité des modalités d'un facteur.

Exemple : reprendre l'étude précédente en indiquant l'état-civil comme question supplémentaire.

3.7 ACM menée à partir d'un tableau disjonctif

3.7.1 L'exemple choisi

L'exemple qui suit est tiré de [Rouanet - Le Roux] qui fait lui-même référence à une célèbre enquête britannique (D.V. Glass, 1954, Social Mobility in Britain, London, Routledge & Kegan Paul).

Une enquête a été menée auprès de 3450 individus. Les variables qui ont été observées sont les suivantes :

- A : statut social du père du répondant (deux modalités : a1 élevé, a2 faible)
- B : niveau d'études du répondant (deux modalités : b1 élevé, b2 faible)
- C : statut du répondant (trois modalités : c1, c2, c3 de niveaux décroissants).

Le tableau des effectifs est donné par :

A	B	C	n_k
a1	b1	c1	106
a1	b1	c2	88
a1	b1	c3	8
a1	b2	c1	18
a1	b2	c2	45

a1	b2	c3	11
a2	b1	c1	106
a2	b1	c2	776
a2	b1	c3	133
a2	b2	c1	27
a2	b2	c2	1274
a2	b2	c3	858
TOTAL			3450

Nous avons en tout 7 modalités, et 2x2x3 = 12 patrons de réponses possibles. Tous sont représentés parmi les réponses observées. Le début du tableau disjonctif complet est le suivant :

	a1	**a2**	**b1**	**b2**	**c1**	**c2**	**c3**
sujet 1	1	0	1	0	1	0	0
sujet 2	1	0	1	0	0	1	0
...	...	0	...	...	...	...	...

Les données observées peuvent également être décrites à l'aide du *tableau disjonctif des patrons* :

	a1	**a2**	**b1**	**b2**	**c1**	**c2**	**c3**	**TOTAL**
a1b1c1	106	0	106	0	106	0	0	**318**
a1b1c2	88	0	88	0	0	88	0	**264**
a1b1c3	8	0	8	0	0	0	8	**24**
a1b2c1	18	0	0	18	18	0	0	**54**
a1b2c2	45	0	0	45	0	45	0	**135**
a1b2c3	11	0	0	11	0	0	11	**33**
a2b1c1	0	106	106	0	106	0	0	**318**
a2b1c2	0	776	776	0	0	776	0	**2328**
a2b1c3	0	133	133	0	0	0	133	**399**
a2b2c1	0	27	0	27	27	0	0	**81**
a2b2c2	0	1274	0	1274	0	1274	0	**3822**
a2b2c3	0	858	0	858	0	0	858	**2574**
TOTAL	**276**	**3174**	**1217**	**2233**	**257**	**2183**	**1010**	**10350**

Enfin, le tableau de Burt est ici :

	A:a1	A:a2	B:b1	B:b2	C:c1	C:c2	C:c3	**Total**
A:a1	276	0	202	74	124	133	19	**828**
A:a2	0	3174	1015	2159	133	2050	991	**9522**
B:b1	202	1015	1217	0	212	864	141	**3651**
B:b2	74	2159	0	2233	45	1319	869	**6699**
C:c1	124	133	212	45	257	0	0	**771**
C:c2	133	2050	864	1319	0	2183	0	**6549**
C:c3	19	991	141	869	0	0	1010	**3030**
Total	**828**	**9522**	**3651**	**6699**	**771**	**6549**	**3030**	**31050**

3.7.2 Lien entre l'ACM et l'AFC

Rappel: 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).

Menée à partir d'un tableau disjonctif, l'ACM est donc une AFC. Et, c'est bien l'onglet "Analyse des correspondances" qui nous servira ici. Grâce à la propriété d'équivalence distributionnelle, il revient au même d'utiliser le tableau disjonctif complet ou le tableau disjonctif des patrons, mais le premier comporte 3450 lignes alors que le second n'en comporte que 12.

Nous pouvons donc, comme en AFC, nous intéresser aux profils ligne et colonne, aux taux de liaison et au Φ^2 d'un des tableaux disjonctifs vu comme un tableau de contingence. On retrouverait les mêmes résultats en effectuant une AFC sur le tableau disjonctif complet.

Rappel : 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 : $\Phi^2 = \frac{7}{3} - 1 = 1,33$.

3.7.3 Valeurs propres

Les fichiers contenus dans le répertoire Statut du serveur de TD permettent de retrouver les résultats qui suivent.

Chargez Statistica et ouvrez le classeur Statut-2007.stw. Utilisez la feuille de données statut-disjonctif-patrons comme feuille active.
Exécutez ensuite une AFC en indiquant que la feuille de données est un tableau de contingence.
Vous devriez obtenir le tableau des valeurs propres (non nulles) suivant :

Valeurs Propres et Inertie de toutes les Dimensions (statut-disjonctif-patrons.sta)					
Table d'Entrée (Lignes x Colonnes) : 12 x 7					
Inertie Totale = 1,3333 Chi2 = 13800, dl = 66 p = 0,0000					
	ValSing.	ValProp.	%age	%age	Chi_
			Inertie	Cumulé	
1	0,7466	0,5574	41,81	41,81	5769,17
2	0,5983	0,3579	26,84	68,65	3704,29
3	0,4824	0,2328	17,46	86,11	2409,25
4	0,4304	0,1852	13,89	100,00	1917,27

3.7.3.1 Calcul des taux d'inertie modifiés

Dans notre exemple, deux valeurs propres sont supérieures à la moyenne (1/3). Les valeurs modifiées conduisent aux résultats suivants :

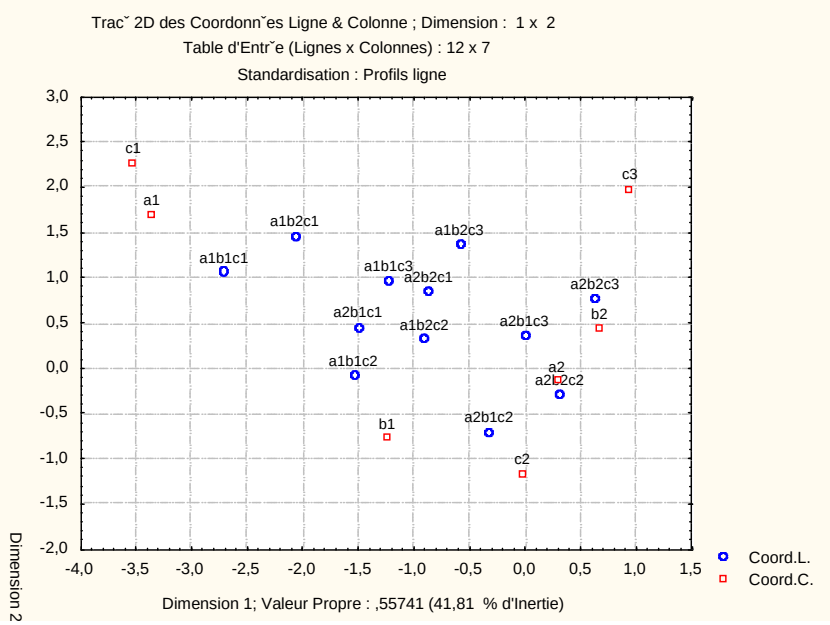
Valeur propre	λ'_l	Taux modifié
0,5574	0,1130	98,81%
0,3579	0,0014	1,19%

On pourrait donc se borner à étudier l'axe 1.

3.7.4 Propriétés géométriques de l'ACM

Les graphiques produits possèdent des propriétés géométriques intéressantes. Cependant, nous avions jusqu'à présent utilisé l'option "Centrer-réduire les données - Profils ligne et colonne" (sous l'onglet "Options" de la fenêtre de dialogue). Or, la mise en évidence de ces propriétés nécessite d'utiliser, selon le cas, l'option "Profils ligne (interpréter dist. lignes)" ou l'option "Profils colonne (interpréter dist. colonnes)".

Par exemple, en utilisant les profils lignes, on obtient :

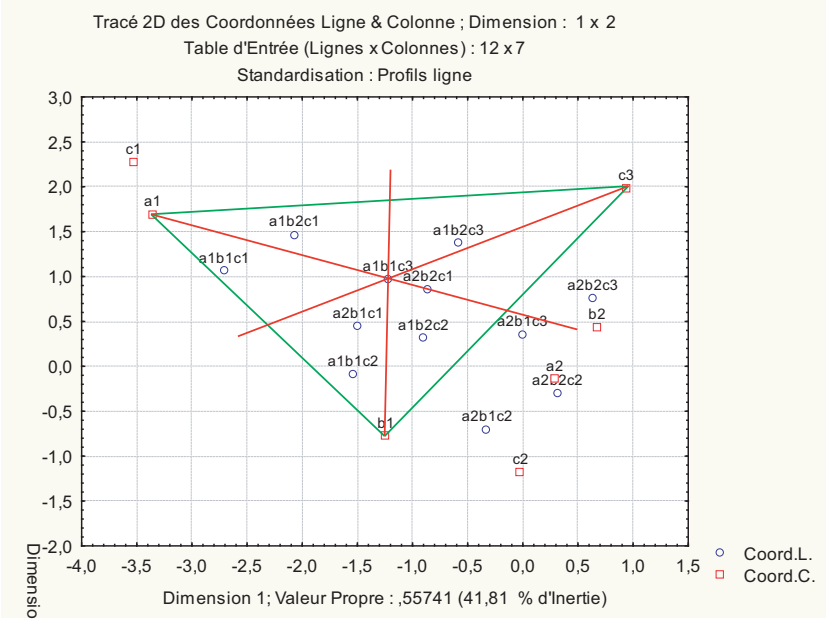


Avec ce choix d'échelles, le nuage des patrons est entièrement contenu à l'intérieur de celui les modalités. Ce graphique met particulièrement bien en évidence les propriétés suivantes :

Le point représentant chaque patron est l'équibarycentre des modalités correspondant à ce patron.

Cette propriété est vraie aussi bien pour les individus que pour les patrons. Elle est vraie dans l'espace multidimensionnel, et elle est conservée par les projections sur les plans factoriels.

De manière plus claire, chaque patron est le centre de gravité du triangle formé par ses trois modalités. Ou encore, par exemple, considérons les modalités a1, b1, c3 et le patron a1b1c3. Chaque droite joignant l'une de ces modalités au point représentant le patron passe par le milieu du segment défini par les deux autres modalités :

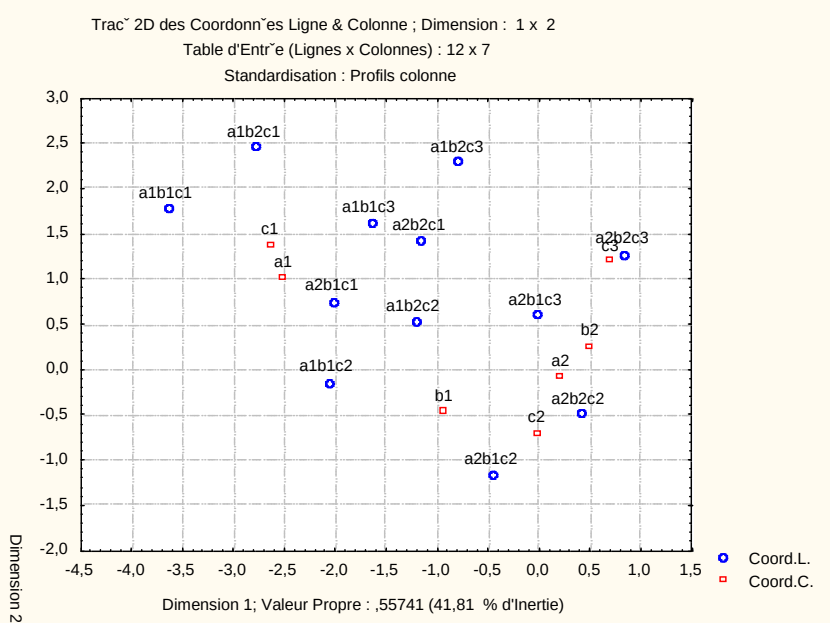


On constate également sur le graphique que les droites (a1 a2) et (b1 b2) passent par l'origine du repère, qui est également à l'intérieur du triangle formé par les trois points c1, c2 et c3.

C'est une conséquence de la propriété suivante :

Le sous-nuage des modalités d'une question a pour point moyen le point moyen du nuage (moyenne pondérée par les fréquences des modalités).

En utilisant les profils colonnes, on obtient :



Chaque modalité est obtenue comme barycentre des patrons qui l'ont choisie, mais chaque patron doit être pondéré par sa fréquence. La constatation la plus immédiate que l'on peut faire sur le graphique ci-dessus est la suivante : chaque modalité se trouve à l'intérieur du polygone (convexe) défini à partir des 4 ou 6 patrons qui l'ont choisie.

On voit également apparaître sur cette représentation la propriété *d'équipollence*, vérifiée par les patrons de réponses. Par exemple les 4 segments suivants :

- le segment qui joint a1b1c1 à a1b1c2
- le segment qui joint a1b2c1 à a1b2c2
- le segment qui joint a2b1c1 à a2b1c2
- le segment qui joint a2b2c1 à a2b2c2

sont parallèles, de même longueur et de même sens

3.7.5 Comparaison entre analyse d'un tableau de Burt et celle d'un TDC

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.

Qu'en est-il avec Statistica ?
Rendez active la feuille de données Statut-Burt.
Effectuez l'analyse en choisissant l'onglet "Analyse des Correspondances Multiple" et l'item "Tableau de Burt".

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. Ce sont également celles que l'on obtiendrait

en effectuant l'analyse à partir de l'onglet "Analyse des Composantes Multiples" et du tableau protocole ou du tableau des effectifs.

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 (statut-Burt.sta) Inertie Totale = ,52730 Chi2 = 16373, dl = 36 p = 0,0000				
	ValSing.	ValProp.	%age	%age	Chi2
			Inertie	Cumulé	
1	0,5574	0,3107	58,9236	58,9236	9647,3580
2	0,3579	0,1281	24,2925	83,2162	3977,3286
3	0,2328	0,0542	10,2761	93,4922	1682,4691
4	0,1852	0,0343	6,5078	100,0000	1065,4910

3.8 Comparaison entre ACM et AFC

Dans le cas où les individus statistiques étudiés sont décrits par deux variables nominales, on peut utiliser une AFC, aussi bien qu'une ACM pour explorer les données recueillies. La question se pose donc d'étudier les relations entre les résultats fournis par ces deux méthodes, dans cette situation.

Nous nous proposons de mener cette étude sur l'exemple "Nobel", déjà utilisé précédemment.

Pour réaliser l'AFC, nous prenons comme données de base le tableau protocole, le tableau d'effectifs ou le tableau de contingence :

PAYS	MEDE	PHYS	CHIM	LITT	SECO
USA	55	43	24	8	9
GB	19	20	21	6	2
RFA	11	14	24	7	0
FRAN	7	9	6	11	0

Pour l'ACM, nous pouvons partir du tableau d'effectifs ou du tableau de Burt :

	USA	GB	RFA	FRAN	MEDE	PHYS	CHIM	LITT	SECO
USA	139	0	0	0	55	43	24	8	9
GB	0	68	0	0	19	20	21	6	2
RFA	0	0	56	0	11	14	24	7	0
FRAN	0	0	0	33	7	9	6	11	0
MEDE	55	19	11	7	92	0	0	0	0
PHYS	43	20	14	9	0	86	0	0	0
CHIM	24	21	24	6	0	0	75	0	0
LITT	8	6	7	11	0	0	0	32	0
SECO	9	2	0	0	0	0	0	0	11

Remarquons les modalités de l'ACM sont formées par la réunion des individus-lignes et des individus colonnes de l'AFC.

L'inertie totale du nuage de points pour l'AFC est le coefficient $\Phi^2 = 0{,}15079$.

Pour l'ACM, l'inertie totale est : $I = \frac{9-2}{2} = 3{,}5$.

3.8.1 Valeurs propres obtenues par l'AFC et par l'ACM

Nombre	Valeurs Propres et Inertie de toutes les Dimensions (Nobel-effectifs.sta)				
de Dims.	Inertie Totale = ,15079 Chi_ = 44,633 dl = 12 p = ,00001				
	ValSing.	ValProp.	%age Inertie	%age Cumulé	Chi2
1	0,320054	λ_1 = 0,102435	67,93261	67,9326	30,32064
2	0,219399	λ_2 = 0,048136	31,92278	99,8554	14,24823
3	0,014767	λ_3 = 0,000218	0,14461	100,0000	0,06455

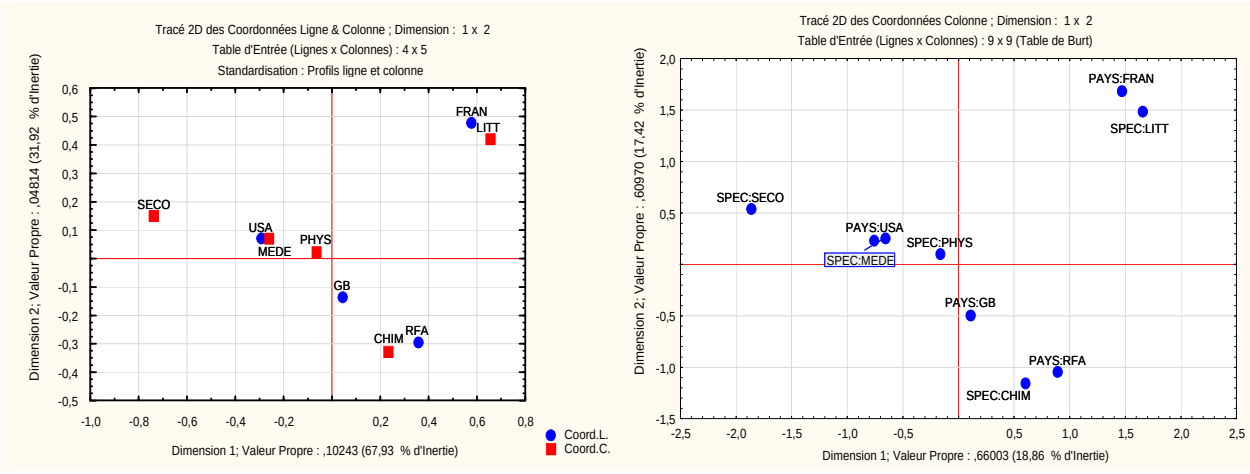
Nombre	Valeurs Propres et Inertie de toutes les Dimensions (Nobel-Burt.sta)				
de Dims.	Inertie Totale = 3,5000				
	ValSing.	ValProp.	%age Inertie	%age Cumulé	Chi2
1	0,812420	μ_1 = 0,660027	18,85792	18,8579	407,5699
2	0,780833	μ_2 = 0,609699	17,41999	36,2779	376,4924
3	0,712309	μ_3 = 0,507383	14,49667	50,7746	313,3117
4	0,707107	μ_4 = 0,500000	14,28571	65,0603	308,7524
5	0,701867	μ_5 = 0,492617	14,07476	79,1350	304,1931
6	0,624740	μ_6 = 0,390301	11,15144	90,2865	241,0124
7	0,583072	μ_7 = 0,339973	9,71351	100,0000	209,9349

Appelons λ_1, λ_2 *et* λ_3 les trois valeurs propres produites par l'AFC. L'ACM produit 7 valeurs propres qui sont reliées aux précédentes par les relations suivantes :

- Pour i = 1, 2, 3, on a : $\mu_i = \frac{1+\sqrt{\lambda_i}}{2}$. Ces valeurs propres sont toutes supérieures à 1/2.
- La 4è valeur propre vaut : $\mu_4 = \frac{1}{2}$. D'une manière générale, le nombre de valeurs propres égales à 1/2 est la différence entre le nombre d'individus lignes et le nombre d'individus colonnes.
- Pour i = 5, 6, 7, on a : $\mu_i = \frac{1-\sqrt{\lambda_{8-i}}}{2}$. Ces valeurs propres sont toutes inférieures à 1/2.

Remarque : A une multiplication par 2 près, le passage des valeurs propres λ_i aux valeurs propres μ_i est la transformation des valeurs propres proposée par Benzécri pour l'ACM. En particulier les pourcentages d'inertie du premier tableau ci-dessus et ceux qui seraient calculés par la méthode de Benzécri sont les mêmes.

3.8.1.1 Coordonnées des points-lignes et points-colonnes de l'AFC et des modalités pour l'ACM



On constate que la représentation graphique des modalités pour l'ACM est analogue à la représentation graphique conjointe des individus-lignes et individus-colonnes. La graduation des axes est cependant différente. En effet, les coordonnées des modalités sur le premier axe, par exemple, sont obtenues en multipliant les coordonnées des individus lignes et des individus colonnes par $\sqrt{\frac{\mu_1}{\lambda_1}}$.

NomLigne	Coordonnées Ligne et Contributions à l'Inertie (Nobel-effectifs.sta)									
	Standardisation : Profils ligne et colonne									
	Ligne	Coord.	Coord.	Masse	Qualité	Inertie	Inertie	Cosinus2	Inertie	Cosinus2
	Numéro	Dim.1	Dim.2			Relative	Dim.1	Dim.1	Dim.2	Dim.2
USA	1	-0,298689	0,070541	0,469595	0,999641	0,293442	0,408993	0,946831	0,048544	0,052810
GB	2	0,042918	-0,134652	0,229730	0,968713	0,031412	0,004131	0,089335	0,086532	0,879378
RFA	3	0,350168	-0,292888	0,189189	0,998635	0,261831	0,226466	0,587570	0,337155	0,411064
FRAN	4	0,575454	0,477359	0,111486	0,999999	0,413315	0,360410	0,592372	0,527768	0,407627

Nom Col.	Coordonnées Colonne et Contributions à l'Inertie (Nobel-effectifs.sta)									
	Standardisation : Profils ligne et colonne									
	Colonne	Coord.	Coord.	Masse	Qualité	Inertie	Inertie	Cosinus2	Inertie	Cosinus2
	Numéro	Dim.1	Dim.2			Relative	Dim.1	Dim.1	Dim.2	Dim.2
MEDE	1	-0,262607	0,071397	0,310811	0,997798	0,152991	0,209248	0,929120	0,032914	0,068678
PHYS	2	-0,069169	0,028409	0,290541	0,915700	0,011765	0,013570	0,783527	0,004871	0,132173
CHIM	3	0,232856	-0,322084	0,253378	0,999783	0,265486	0,134121	0,343189	0,546057	0,656593
LITT	4	0,649223	0,421200	0,108108	0,999914	0,429420	0,444835	0,703713	0,398444	0,296200
SECO	5	-0,739184	0,151474	0,037162	0,999839	0,140337	0,198226	0,959545	0,017714	0,040294

NomLigne	Coordonnées Colonne et Contributions à l'Inertie (Nobel-Burt.sta)									
	Inertie Totale = 3,5000									
	Ligne	Coord.	Coord.	Masse	Qualité	Inertie	Inertie	Cosinus2	Inertie	Cosinus2
	Numéro	Dim.1	Dim.2			Relative	Dim.1	Dim.1	Dim.2	Dim.2
PAYS:USA	1	-0,75819	0,25105	0,234797	0,564745	0,075772	0,204496	0,508943	0,024272	0,055802
PAYS:GB	2	0,10894	-0,47922	0,114865	0,072033	0,110039	0,002065	0,003540	0,043266	0,068493
PAYS:RFA	3	0,88886	-1,04238	0,094595	0,437879	0,115830	0,113233	0,184351	0,168578	0,253528
PAYS:FRAN	4	1,46072	1,69890	0,055743	0,629884	0,126931	0,180205	0,267729	0,263884	0,362156
DISC:MEDE	5	-0,66660	0,25410	0,155405	0,229512	0,098456	0,104624	0,200394	0,016457	0,029118
DISC:PHYS	6	-0,17558	0,10111	0,145270	0,016811	0,101351	0,006785	0,012624	0,002436	0,004186
DISC:CHIM	7	0,59108	-1,14628	0,126689	0,564482	0,106660	0,067061	0,118566	0,273029	0,445916
DISC:LITT	8	1,64798	1,49904	0,054054	0,601569	0,127413	0,222418	0,329192	0,199222	0,272377
DISC:SECO	9	-1,87633	0,53909	0,018581	0,147101	0,137548	0,099113	0,135884	0,008857	0,011217

On peut enfin remarquer que, si l'on réalise une AFC en indiquant Nobel-Burt.sta comme tableau de contingence, Statistica produit des graphiques analogues aux précédents, mais indique encore des valeurs propres et une inertie totale différentes. Les valeurs propres sont alors les carrés des valeurs propres indiquées pour l'ACM.

3.9 Exercices

3.9.1 Exercice 1

Les fichiers Beverage.sta et Beverag2.sta sont des fichiers d'exemples fournis avec Statistica pour illustrer l'ACM. Ils ont été recopiés dans le répertoire Boissons du serveur de TD.

Le fichier d'exemple Beverage.sta contient des données collectées sur un groupe d'étudiants en maîtrise de gestion, hommes et femmes, de l'Université de Columbia, auxquels on a demandé d'indiquer la fréquence avec laquelle ils avaient acheté et consommé différents types de soda durant du mois écoulé. Les données pour les 34 individus ont été codifiées dans un tableau disjonctif complet (binaire) : un 1 a été saisi si l'individu a répondu avoir acheté ou consommé au moins une fois au cours du mois la boisson respective, et un 0 a été saisi si l'individu respectif a répondu avoir acheté ou consommé moins d'une fois dans le mois. Pour chacun des 8 sodas populaires utilisés dans cette étude, une seconde variable a été créée, codifiée comme l'inverse de la première variable respective, c'est-à-dire qu'un 1 a été saisi si la boisson respective n'a pas été consommée ni achetée, et 0 a été saisi si elle a été consommée ou achetée au cours du mois. Ci-dessous, observez une liste partielle des données codifiées de cette manière, pour 8 sodas courants. Ouvrez le fichier de données Beverage.sta situé dans le répertoire Boissons.

1) Réalisez une ACM sur ces données et retrouver ainsi les principales conclusions données dans l'aide de Statistica :

Il s'avère que toutes les boissons sont raisonnablement bien représentées par la solution à deux dimensions, mis à part Pepsi Light dont la valeur de Qualité est inférieure à 0,5.
Un examen attentif du graphique suggère que le premier axe oppose essentiellement les boissons allégées aux boissons classiques, alors que la seconde dimension oppose les colas aux autres sodas.
En outre, si vous examinez attentivement les statistiques des coordonnées de lignes, vous allez constater que les individus contribuant le plus à l'inertie de la seconde dimension sont les observations numéro 13 et 28. Ces points "définissent" presque à eux seuls la direction de la seconde dimension.

2) Reprenez ensuite l'étude à l'aide du fichier Beverag2.sta, dont les données sont saisies sous forme de tableau protocole.

3.9.2 Exercice 2

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 une feuille de données croisant la fréquentation (variable 1) et l'âge (variable 13) sous la forme d'un tableau de contingence et cinq feuilles de données contenant les tableaux de Burt obtenus en sélectionnant 3 ou 4 des items du questionnaire. Choisissez l'un de ces tableaux de Burt et analysez-le à l'aide d'une ACM.

3.9.3 Exercice 3

Le fichier Enquete-Eleves-US.stw contient les données d'une étude trouvée (en 2006) sur le site http://stat.genopole.cnrs.fr/teaching/. Ces pages semblent malheureusement ne plus être accessibles en 2007.
Il s'agit d'une étude sur des élèves provenant de différentes écoles du milieu rural au milieu urbain. Les différentes questions et leurs modalités sont :

Gender: Boy or girl
Grade: 4, 5 or 6
Age: Age in years
Race: White, Other
Urban/Rural: Rural, Suburban, or Urban school district
School: Brentwood Elementary, Brentwood Middle, Ridge, Sand, Eureka, Brown, Main, Portage, Westdale Middle"
Goals: Student's choice in the personal goals question where options were 1 = Make Good Grades, 2 = Be Popular, 3 = Be Good in Sports.
Grades: Rank of "make good grades" (1=most important for popularity, 4=least important)
Sports: Rank of "being good at sports" (1=most important for popularity, 4=least important)
Looks: Rank of "being handsome or pretty" (1=most important for popularity, 4=least important)
Money: Rank of "having lots of money" (1=most important for popularity, 4=least important)

A partir d'un questionnement de votre choix sur ces données, faites une sélection de variables (au moins 4 variables) et procédez à une ACM en utilisant ces variables.
Interprétez ensuite (brièvement) des résultats obtenus.

3.9.4 Exercice 4

Source : Kolié, Ouo-Ouo Jean-Philippe. (2009). Identification des groupes homogènes de maraîchers et l'évaluation de leurs performances économiques au Burkina Faso. Montpellier : CIHEAM-IAMM. 115p. (Master of Science ; n° 101).

Le travail cité *supra* vise à identifier les différentes catégories de maraîchers burkinabés et à évaluer leurs performances en matière de production. Pour ce faire, l'auteur a eu recours à une analyse des correspondances multiples, complétée par une typologie. L'analyse s'est faite sur la base de données du Ministère de l'Agriculture du Burkina Faso.

On reprend ici une partie de l'étude menée en considérant cinq des variables utilisées par l'auteur :

- Le genre ; deux modalités : M, F.
- La classe d'âge ; sept modalités : 20-25 ans, 25-30 ans, 30-35 ans, 35-40 ans, 40-50 ans, > 50 ans, < 20 ans.
- L'encadrement : un site encadré est un site dont les maraîchers reçoivent ou ont reçu des conseils sur les techniques culturales données par un encadreur institutionnel, un membre d'une ONG ou d'un projet, etc.; deux modalités : Non, Oui.
- L'autoconsommation ; cinq modalités : de 10% à 20%, plus de 20%, de 5% à 10%, 0% (sans autoconsommation), moins de 5%.
- Le revenu, en milliers de francs CFA (1000 FCFA = 1,52 €) ; six modalités : de 15 à 30 milliers de FCFA, de 30 à 45 milliers de FCFA, de 45 à 65 milliers de FCFA, de 5 à 15 milliers de FCFA, moins de 5 milliers de FCFA, plus de 65 milliers de FCFA

Les données traitées portent sur 863 maraîchers. Elles sont résumées dans la feuille de données du fichier Maraichers.stw.

1) a) Sous quelle forme les données observées sont-elles présentées ? Quel nom donne-t-on à ce type de tableau ?
b) A l'intersection de la ligne et de la colonne étiquetées "AGE:20-25", on trouve le nombre 116. De manière analogue, on trouve le nombre 69 à l'intersection de la ligne étiquetée "AGE:20-25" et de la colonne étiquetée "Enc.:N".
Quelles significations peut-on donner à ces valeurs ?

2) Quelle méthode d'analyse des données multidimensionnelles peut-on utiliser ici ? A quel type de données cette méthode s'applique-t-elle généralement ?

3) a) L'inertie totale du nuage de points est $I = 3{,}4$. Comment peut-on retrouver cette valeur ?
On sait que l'inertie relative des questions dépend seulement de leur nombre de modalités. Calculer l'inertie relative de chacune des six questions.

b) Pour la méthode utilisée, quelles sont les recommandations généralement indiquées en ce qui concerne le nombre de modalités des différentes questions et les fréquences des modalités ? Dans quelle mesure ces conditions sont-elles vérifiées sur l'exemple traité ici ?

4) Commenter et interpréter le tableau des valeurs propres. Construire le tableau des taux d'inertie modifiés. Justifier le choix de n'étudier en détail que les deux premiers axes. :

5) Etude des deux premiers axes.

a) Quelles sont les modalités dont la contribution est supérieure à la moyenne sur le premier axe ? Pour chacune d'elles, préciser le signe de la coordonnée correspondante. Comment peut-on interpréter cet axe en termes d'opposition entre modalités ?

b) Même question pour le deuxième axe.

6) a) L'une des questions a très peu contribué à l'inertie du premier axe. Laquelle ?

b) Pour les questions comportant 2 modalités (Genre, Encadrement), on voit que chacune des deux modalités se voit attribuer le même Cosinus^2. Ce résultat est-il étonnant ? Pourquoi ?

3.9.5 Exercice 5

Source : Hmam, R., Galmiche, J., Analyse d'une enquête : l'intégration de la communauté du sous-continent indien à Northampton, Les Cahiers de l'Analyse des Données, Vol. XX, 1995, n° 4, pp. 413-432

Dans le cadre de sa thèse, l'un des auteurs (R. Hmam) a mené une enquête par questionnaire sur l'intégration des immigrés originaires du sous-continent indien à la société britannique. L'enquête a été menée dans la ville de Northampton. L'échantillon interrogé est de 384 personnes et a été constitué en équilibrant sensiblement les trois nationalités (Indiens, Bengalais, Pakistanais), alors que les Pakistanais sont nettement moins nombreux que les deux autres communautés dans la ville. On s'intéresse ici à la partie "signalement" du questionnaire : sexe, âge, religion, nationalité, etc.

Six questions sont prises en compte :

- Le sexe ; deux modalités : M, F.
- La classe d'âge ; trois modalités : de 15 à 25 ans (age1), de 25 à 40 ans (age2), 40 ans et plus (age3).
- La religion ; trois modalités : hindou, sikh, musulman.
- La nationalité ; trois modalités : Inde, Bengale, Pakistan.

- Le lieu de naissance ; trois modalités : en Grande-Bretagne, dans le sous-continent indien, ailleurs.
- La personne interrogée a-t-elle des enfants ? Deux modalités : oui, non.

On notera que deux questionnaires ont été écartés de la présente étude, qui porte donc sur 382 réponses. Les données observées pour ces 6 variables sont résumées dans la feuille de données "Signalement" du classeur Statistica Enquete-Northampton.stw :

1) a) Sous quelle forme les données observées sont-elles présentées ? Quel nom donne-t-on à ce type de tableau ?
b) A l'intersection de la ligne et de la colonne étiquetées "age2", on trouve le nombre 107. De manière analogue, on trouve le nombre 33 à l'intersection de la ligne étiquetée "age2" et de la colonne étiquetée "Inde". Quelles significations peut-on donner à ces valeurs ?

2) Quelle méthode d'analyse des données multidimensionnelles peut-on utiliser ici ? Traiter à l'aide de cette méthode, les données proposées.

3) a) L'inertie totale du nuage de points est I = 1,666. Comment peut-on retrouver cette valeur ?
On sait que l'inertie relative des questions dépend seulement de leur nombre de modalités. Calculer l'inertie relative des questions comportant 2 modalités, comportant 3 modalités.
b) Pour la méthode utilisée, quelles sont les recommandations généralement indiquées en ce qui concerne le nombre de modalités des différentes questions et les fréquences des modalités ? Dans quelle mesure ces conditions sont-elles vérifiées sur l'exemple traité ici ?

4) Commenter et interpréter le tableau des valeurs propres. Déterminer les taux d'inertie modifiés et justifier le choix de n'étudier en détail que les deux premiers axes.

5) Etude des deux premiers axes.
a) Quelles sont les modalités dont la contribution est supérieure à la moyenne sur le premier axe ? Pour chacune d'elles, préciser le signe de la coordonnée correspondante. Comment peut-on interpréter cet axe en termes d'opposition entre modalités ?
b) Même question pour le deuxième axe.

6) a) L'une des questions a très peu contribué à l'inertie des deux premiers axes. Laquelle ? Que peut-on en déduire concernant les liens éventuels entre les réponses à cette question et les réponses aux autres questions ?
b) Commenter les qualités de représentation des modalités de cette question dans le premier plan factoriel. Ce résultat est-il étonnant ? Pourquoi ?

7) a) Quelle est la modalité qui a le plus influé sur la formation de l'axe 3 ? En quoi les réponses des sujets ayant choisi cette modalité se distinguent-elles des autres réponses ? Comment peut-on résumer cet axe ?
b) Quelle est la question qui a le plus influencé la formation de l'axe 4 ? Comment peut-on résumer cet axe ?

3.10 Exercice à rendre

Ref. Cottet M., Piégay H., Honegger A., Modélisation des préférences esthétiques : vers la prise en compte des perceptions dans les projets de restauration écologique de bras morts, Neuvièmes rencontres de Théo Quant., Besançon, 4-6 mars 2009, http://thema.univ-fcomte.fr

Afin de pallier la dégradation des écosystèmes, les projets de restauration écologique se multiplient depuis les années 1990. Caractériser et anticiper la perception du public sont deux enjeux majeurs afin de parvenir à une gestion durable des écosystèmes qui soit partagée par les acteurs. Tel est le but de cette recherche, appliquée à un écosystème particulier, celui des bras morts bordant le Rhône et l'un de ses affluents, l'Ain dans sa basse vallée.

L'objectif de cette étude est d'expliquer la perception esthétique d'un plan d'eau donné (variable dépendante correspondant à une note d'esthétique attribuée par des individus) par une série de variables explicatives se rapportant à certains attributs physiques de plans d'eau (variables indépendantes).

La variable dépendante : une évaluation esthétique des plans d'eau de bras morts
Une enquête de perception esthétique de plans d'eau d'anciens bras a été réalisée. Elle repose sur l'utilisation d'un photo-questionnaire. Selon cette technique d'enquête, il est demandé à des répondants de réagir à une série de photographies qui leur est présentée. Les réponses ont été recueillies auprès d'une population expérimentale de 100 étudiants en licence de géographie. Trente-quatre plans d'eau (désignés par des codes à une ou deux lettres (A, AA, ..., O) ont ainsi été évalués. La note d'esthétique (sur 10) attribuée à chaque plan d'eau est la moyenne des évaluations faites par les sujets.

Les variables explicatives : un jeu d'attributs visuels caractérisant les plans d'eau de bras morts
Une sélection des variables les plus pertinentes pour expliquer les préférences esthétiques en matière de plans d'eau de bras morts a été réalisée. Six variables semblent avoir une influence tranchée sur les jugements esthétiques. La dominance de vert ainsi que la présence de couleurs chaudes et vives semblent avoir une influence positive sur les perceptions, tandis que la dominance de gris/marron, une eau trouble, la présence d'algues aux formes mal structurées ainsi que la présence de sédiments paraissent avoir un impact négatif. Ces six variables dichotomiques ont alors été sélectionnées pour construire le modèle des préférences esthétiques de plans d'eau de bras morts.
Les 12 premières colonnes de la feuille de données présente dans le classeur Perception-esthetique.stw donnent les valeurs de ces 6 variables sur les 34 photographies.

Ces données ont été traitées par une méthode d'analyse multidimensionnelle.

1) a) Observer le tableau des données fournies. Quel nom donne-t-on à un tel tableau ?

b) On applique une analyse factorielle des correspondances (AFC) à ce tableau, considéré comme un tableau de contingence. Quel autre nom donne-t-on encore à cette méthode ?

2) a) Le tableau 2 donne les valeurs propres produites par l'AFC. On observe que seules les 6 premières sont non nulles. Ce résultat est-il étonnant ? Pourquoi ?

b) En raisonnant sur les seules valeurs propres non nulles, déterminer le nombre d'axes factoriels qu'il semble pertinent d'étudier.

3) Etude des individus colonnes (modalités des variables)

a) Pour le premier axe factoriel, quels sont les individus colonnes dont la contribution est supérieure à la moyenne ? Pour chacun d'eux, préciser le signe de la coordonnée correspondante.

b) Même question pour le deuxième axe.

c) Les auteurs caractérisent les deux axes en termes de couleur dominante et de présence éventuelle d'objets environnementaux (végétation, sédiments). Caractériser les deux axes selon ces critères.

4) Examiner le graphique des individus-lignes. Expliquer pourquoi les 34 individus y sont représentés par seulement 15 points distincts.

5) On a construit ensuite un modèle de régression linéaire (régression factorielle) en utilisant la note d'esthétique (variable Note esthétique) comme variable dépendante et les coordonnées sur les deux premiers axes factoriels comme prédicteurs.

Calculer le coefficient de corrélation multiple et donner l'équation de régression.

Travail à rendre par mail à votre enseignant (Francois.Carpentier@univ-brest.fr) :

- Un classeur Statistica contenant les résultats numériques des analyses et les graphiques.
- Un fichier Word ou LibreOffice Writer ou un rapport Statistica fournissant l'interprétation des résultats.