4 Analyse des Correspondances Multiples

4.1 Introduction

L'analyse factorielle des correspondances, vue dans le paragraphe précédent, s'applique à des situations où les individus statistiques sont décrits par *deux* variables nominales. Mais il est fréquent que l'on dispose d'individus décrits par *plusieurs* (deux ou plus) variables nominales ou ordinales. C'est notamment le cas lorsque nos données sont les résultats d'une enquête basée sur des questions fermées à choix multiples. 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.

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.

4.2 Exemple

4.2.1 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 vivez-vous ? (variable "éloignement")	17	moins de 1 km	26,89%	< 1 km
	18	1 à 5 km	49,87%	1 à 5 km
	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

4.2.2 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).

4.2.2.1 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
...					

4.2.2.2 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

i 4	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1
i 5	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0
...	...	...																				

4.2.2.3 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 1km	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
...	...	...																				

4.2.2.4 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 1km	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.

4.3 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. Mais, ce tableau comporte 383 lignes. Cependant, 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.

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

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

Le profil ligne moyen 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 :

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

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 $f_{\bullet k}$ désignent des quantités différentes : f_k est la fréquence de la modalité k dans la population étudiée; $f_{\bullet k}$ est définie comme pour l'AFC, fréquence ligne marginale de la k-ième colonne du tableau disjonctif des patrons.

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

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. 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	Coloc	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.

4.3.1.1 Distance d'un patron au profil 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 :

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

4.3.1.2 Distance entre deux patrons

De même, la distance entre deux 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 :

$$d^2_{\Phi^2}(L_i, L_{i'}) = \frac{1}{Q} \sum_k \frac{(\delta_{ik} - \delta_{i'k})^2}{f_k}$$

Notations utilisées : Li et Li' désignent deux patrons, Q est le nombre de questions. δ_{ik} prend la valeur 1 si la modalité k fait partie du patron i, et la valeur 0 sinon. Enfin, f_k est la fréquence de la modalité k dans la population.

Signification de l'expression $(\delta_{ik} - \delta_{i'k})^2$: cette expression est égale à 0 si la modalité k fait partie des deux patrons, ou ne fait partie d'aucun d'entre eux, elle est égale à 1 si la modalité fait partie d'un seul des deux patrons. Ainsi, on calcule la somme des inverses des fréquences des modalités qui font partie de l'un ou l'autre patron, sans faire partie des deux patrons, puis on divise cette somme par le nombre de questions.

Cette formule montre que deux individus (ou deux patrons) sont d'autant plus éloignés que leurs réponses diffèrent pour un plus grand nombre de questions et pour des modalités rares. Cette formule peut encore être écrite sous la forme :

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

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 :

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

4.3.2 Distances entre modalités (profils colonnes du tableau disjonctif)

4.3.2.1 Distance d'une modalité au profil colonne moyen

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

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

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.

4.3.2.2 Distance entre deux modalités

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.

4.4 Inertie du nuage de points. Contributions

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

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

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 :

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

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 :

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

4.5 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-lignes (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 variable est l'origine des axes.

4.5.1 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

4.5.1.1 Taux modifiés proposés par Benzécri

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.

4.1.2 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.

4.1.3 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.

4.1.3.1 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".

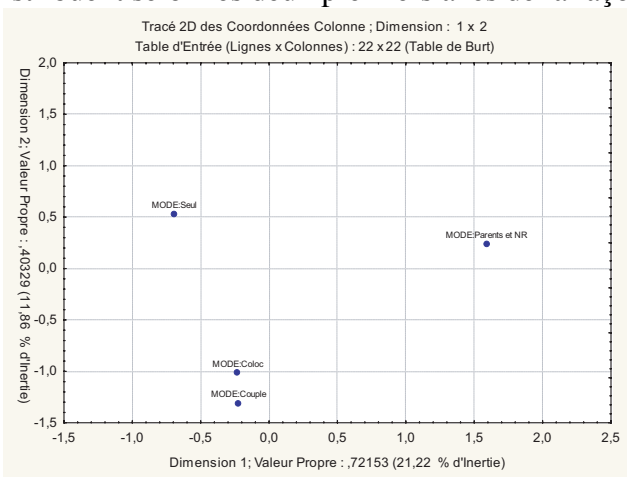
Constitution du tableau précédent :

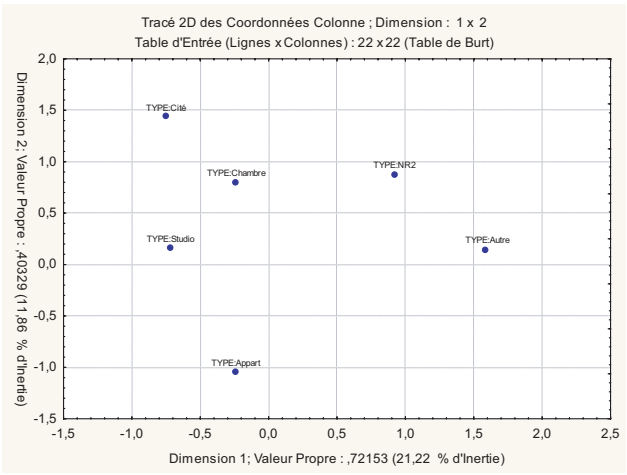
Dans le tableau des résultats relatifs aux modalités, on additionne les inerties des différentes modalités d'une même variable, pour le nuage entier d'une part, et pour chacun des axes retenus d'autre part. Par exemple, pour la première variable :

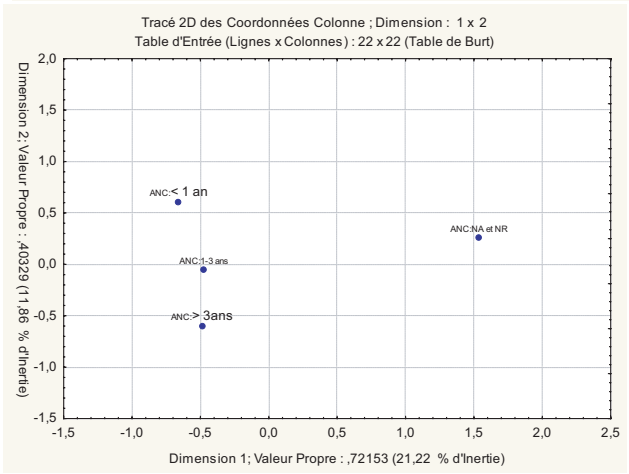
	Inertie	Inertie dim 1	Inertie dim 2	Inertie dim 3
MODE:Seul	0,0304	0,0641	0,0661	0,0170
MODE:Coloc	0,0507	0,0020	0,0714	0,0521
MODE:Couple	0,0511	0,0018	0,1120	0,0048
MODE:Parents et NR	0,0442	0,1741	0,0070	0,0015
Total	**0,1764**	**0,2420**	**0,2565**	**0,0754**

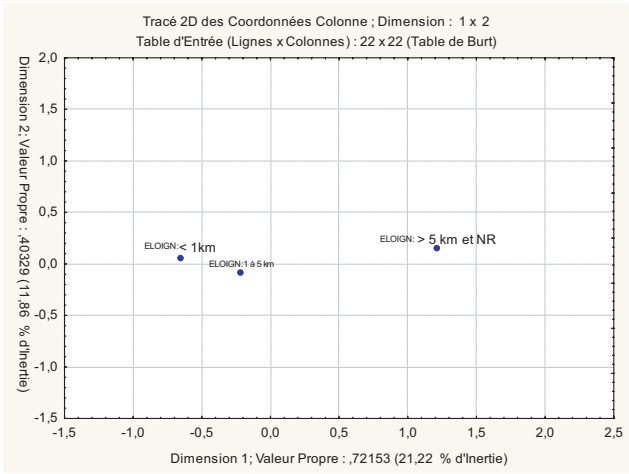
Graphiques "par variable"

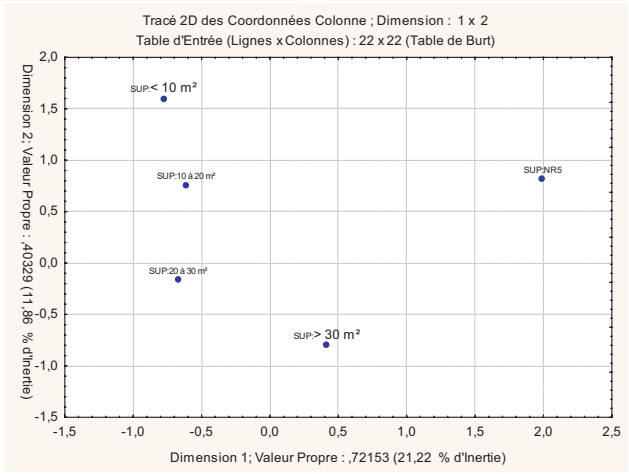
Dans certains cas, il peut être intéressant de réaliser des graphiques montrant la disposition des modalités d'une variable par rapport à 2 axes factoriels. Pour notre exemple, les modalités de chacune des variables se distribuent selon les deux premiers axes de la façon suivante :











4.1.3.2 Etude des axes

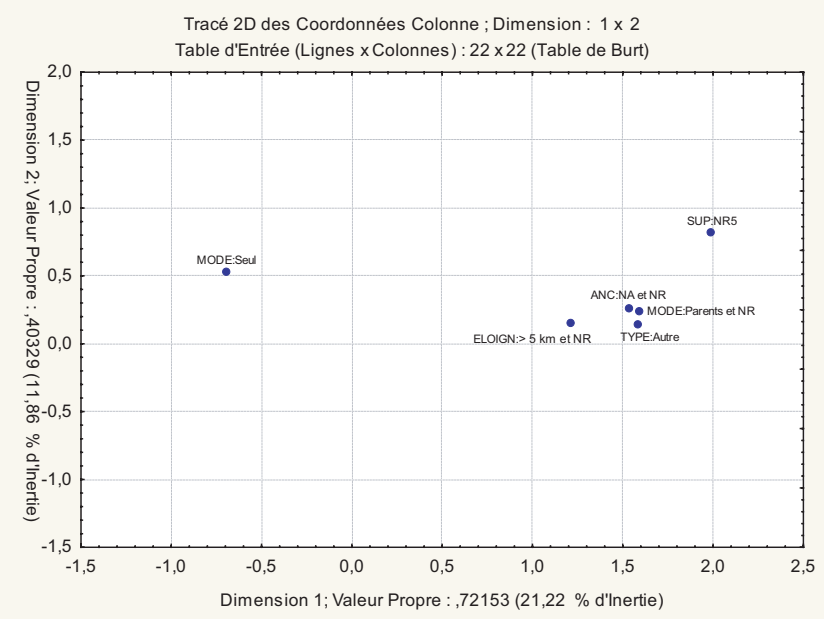
Pour chacun des axes, on pourra repérer les modalités dont la contribution à la formation de l'axe est supérieure à la moyenne (ici, 1/22=0,045), et éventuellement réaliser un graphique limité à ces seules modalités.

Premier Axe

Ainsi, pour le premier axe, on obtient :

-	+
MODE: Seul (6,41%)	MODE: Parents / NR4 (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 :

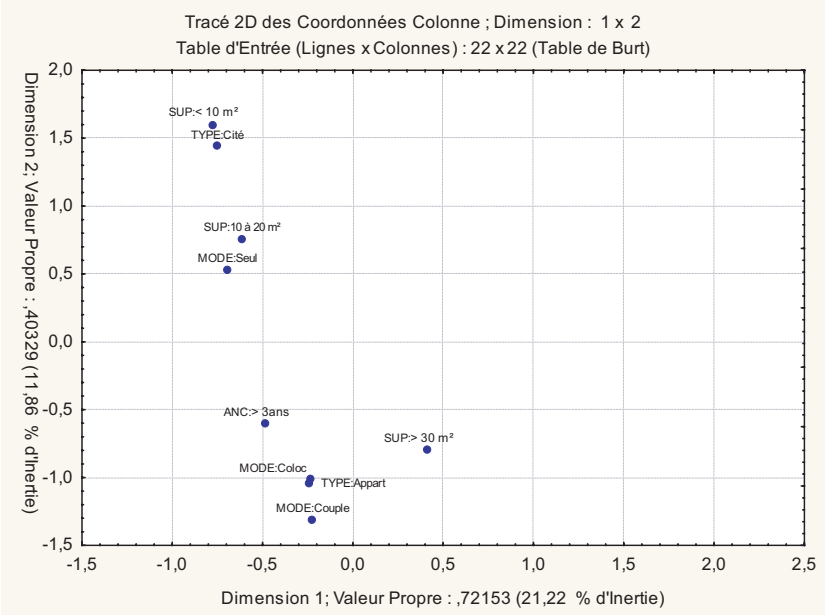


On voit que 4 des 5 variables figurent dans ce tableau, mais que les modalités concernées, dans la partie positive de l'axe sont essentiellement NA, NR, Autre et "Parents" pour la variable "MODE". A l'opposé, avec une contribution plus modeste, on trouve la modalité "Seul" de la variable "MODE". Cet 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%)

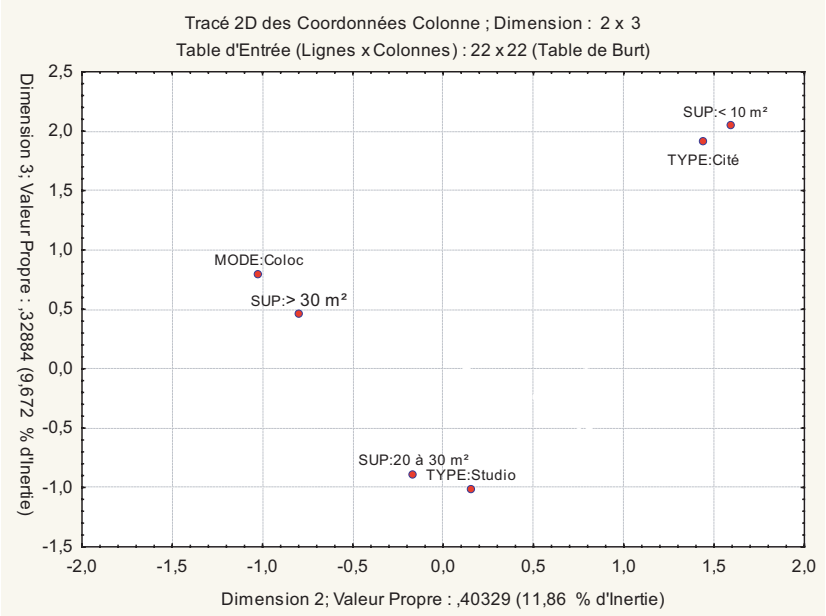


Cet axe fait essentiellement intervenir les variables TYPE, SUPERFICIE et MODE. Les modalités mises en jeu concernent essentiellement les étudiants qui n'habitent plus au foyer familial. L'axe oppose clairement les étudiants vivant seuls en cité universitaire, dans un logement de faible superficie (partie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

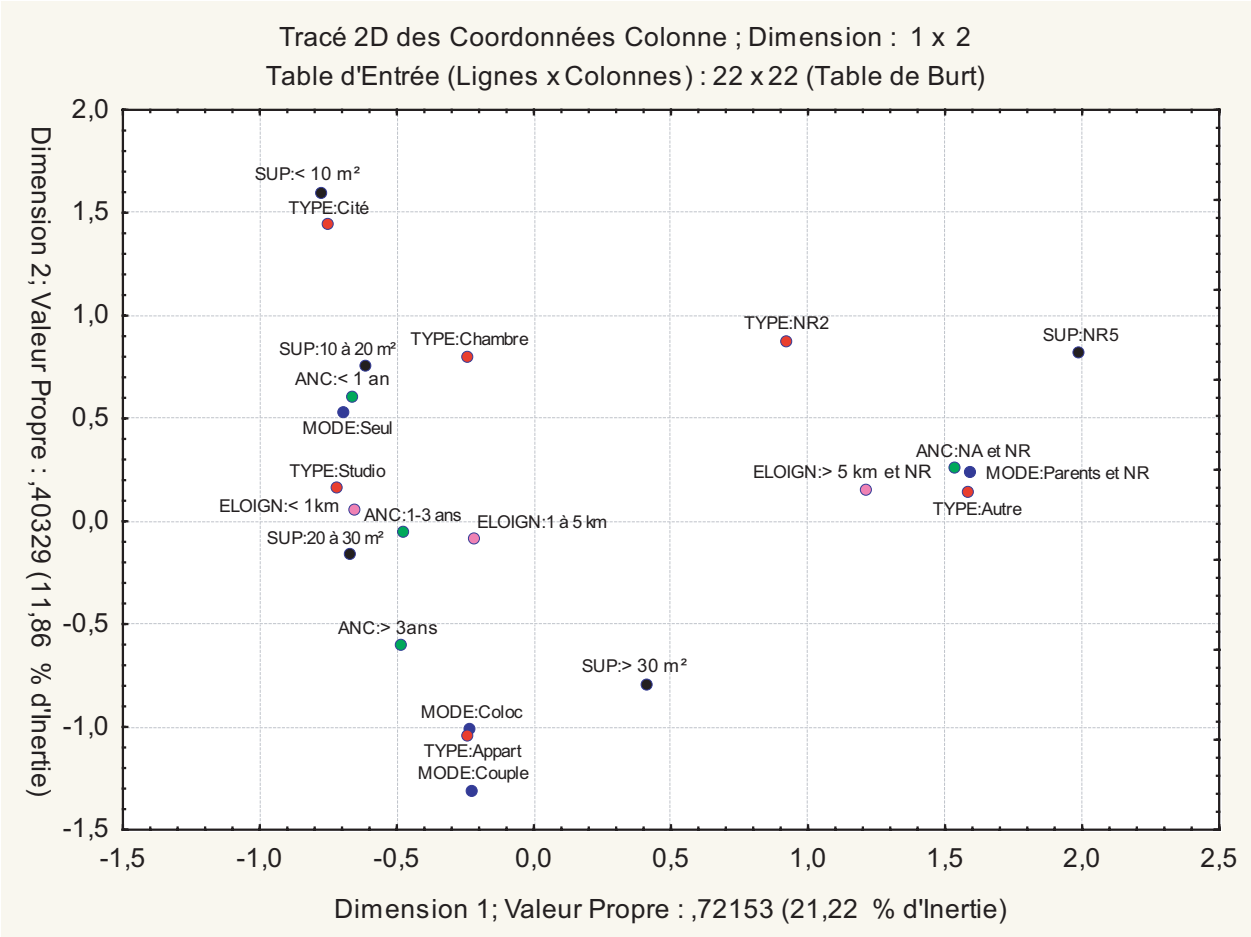
Pour le 3ème axe, on obtient :

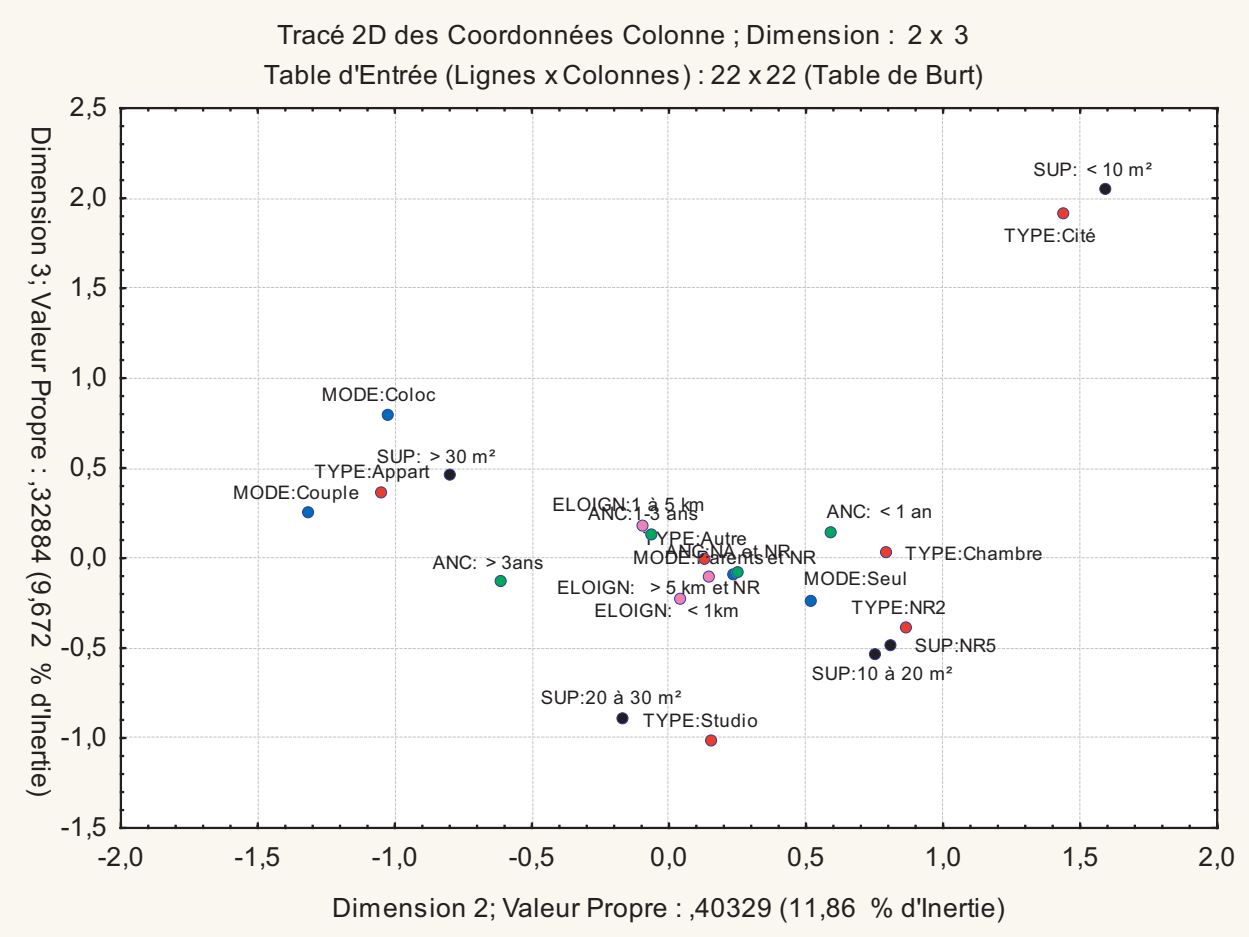
-	+
TYPE: Studio (18,0%) SUPER: 20 à 30m^2 (12,1%)	TYPE: Cité (23,8%) SUPER: <10m^2 (23,3%) MODE: Coloc (5,2%) SUPER: >30m^2 (4,85%)



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.

4.1.3.3 Graphiques selon les deux premiers plans principaux

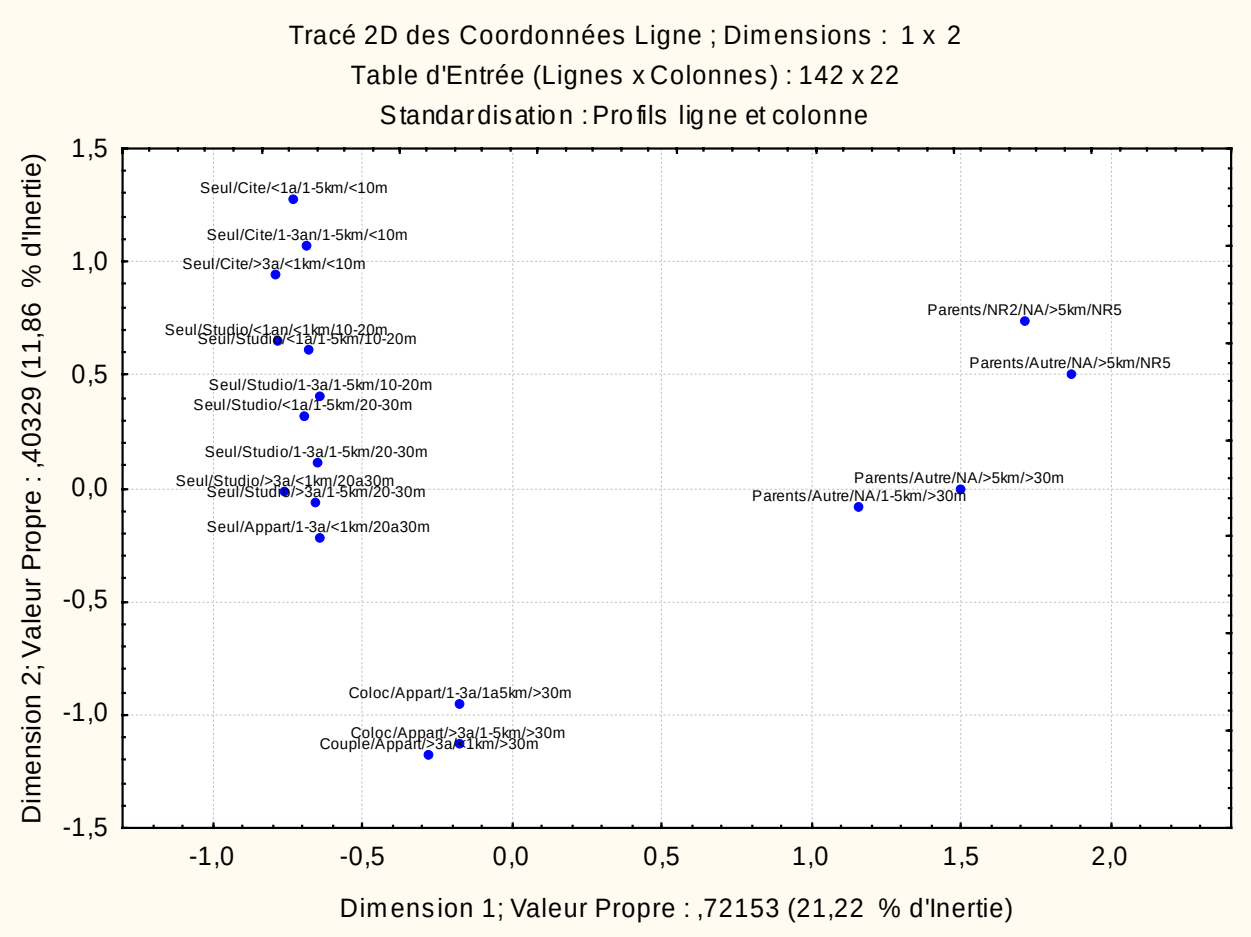




L'étude précédente a permis de distinguer essentiellement trois groupes : le groupe des étudiants logeant chez leurs parents, et des non réponses (partie positive du premier axe), le groupe des étudiants vivant seuls en cité ou en chambre (partie positive du deuxième axe) et le groupe des étudiants vivant 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.

4.1.3.4 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 :


Tracé 2D des Coordonnées Ligne ; Dimensions : 1 x 2
Table d'Entrée (Lignes x Colonnes) : 142 x 22
Standardisation : Profils ligne et colonne
Dimension 2; Valeur Propre : ,40329 (11,86 % d'Inertie)
Dimension 1; Valeur Propre : ,72153 (21,22 % d'Inertie)
1,5
1,0
0,5
0,0
-0,5
-1,0
-1,5
-1,0
-0,5
0,0
0,5
1,0
1,5
2,0
Seul/Cite/<1a/1-5km/<10m
Seul/Cite/1-3an/1-5km/<10m
Seul/Cite/>3a/<1km/<10m
Seul/Studio/<1an/<1km/10-20m
Seul/Studio/<1a/1-5km/10-20m
Seul/Studio/1-3a/1-5km/10-20m
Seul/Studio/<1a/1-5km/20-30m
Seul/Studio/1-3a/1-5km/20-30m
Seul/Studio/>3a/<1km/20a30m
Seul/Studio/>3a/1-5km/20-30m
Seul/Appart/1-3a/<1km/20a30m
Coloc/Appart/1-3a/1a5km/>30m
Coloc/Appart/>3a/1-5km/>30m
Couple/Appart/>3a/<1km/>30m
Parents/NR2/NA/>5km/NR5
Parents/Autre/NA/>5km/NR5
Parents/Autre/NA/>5km/>30m
Parents/Autre/NA/1-5km/>30m