ANALYSE FACTORIELLE DES CORRESPONDANCES

Sandrine Mignon-Grasteau

2014

AFC : Quand et pourquoi ?

- Pour l'analyse de données **qualitatives** (par opposition aux données quantitatives analysées par ACP)

- Les **tableaux** de contingence analysés sont **différents de ceux de l'ACP** :
  - Lignes : modalités d'une variable discontinue
  - Colonnes : modalités d'une autre variable discontinue
  - Cellules : Fréquence des deux modalités conjointes

AFC : importance du χ^2

- Donne une première indication sur l'existence d'une dépendance entre variables
- Ex : La préférence pour la bière évolue-t-elle avec l'âge ?

	21-30 ans	31-40 ans	41-50 ans	Totaux
Goût	3	1	8	12
Couleur	3	4	2	9
Alcool	4	5	0	9
Totaux	10	10	10	30

AFC : importance du χ^2

- Effectif observé

	21-30 ans	31-40 ans	41-50 ans	Totaux
Goût	3	1	8	12
Couleur	3	4	2	9
Alcool	4	5	0	9
Totaux	10	10	10	30

AFC : importance du χ^2

- Effectif théorique si indépendance

$$\text{Effectif théorique} = \frac{\text{Nb d'individus dans la ligne} \times \text{Nb d'individus dans la colonne}}{\text{Nb d'individus dans le tableau}}$$

donc, par exemple

$$\text{Effectif théorique(Goût, 21-30 ans)} = \frac{12 \times 10}{30} = 4$$

$$\text{Effectif théorique(Alcool, 41-50 ans)} = \frac{10 \times 9}{30} = 3$$

	21-30 ans	31-40 ans	41-50 ans	Totaux
Goût	4	4	4	12
Couleur	3	3	3	9
Alcool	3	3	3	9
Totaux	10	10	10	30

AFC : importance du χ^2

- Ecart à l'indépendance

	21-30 ans	31-40 ans	41-50 ans
Goût	-1	-3	4
Couleur	0	1	-1
Alcool	1	2	-3

- Les 41-50 ans ont des écarts à l'indépendance plus élevés que les deux autres groupes
  - ↳ Ils ont des préférences plus marquées
- Les 41-50 ans ont des écarts à l'indépendance de signe opposé à ceux des autres groupes
  - ↳ Leurs goûts sont différents de ceux des autres groupes

AFC : importance du χ^2

- Carrés des écarts à l'indépendance / effectif théorique

	21-30 ans	31-40 ans	41-50 ans
Goût	1/4	9/4	16/4
Couleur	0/3	1/3	1/3
Alcool	1/3	4/3	9/3

$$\chi^2 = \sum \frac{(\text{Effectif observé} - \text{Effectif théorique})^2}{\text{Effectif théorique}} = 11.83$$

Condition : moins de 20% des cellules ont un effectif < 5

AFC : importance du χ^2

- Interprétation des sommes par ligne et colonne

	21-30 ans	31-40 ans	41-50 ans	Totaux
Goût	1/4	9/4	16/4	6.50
Couleur	0/3	1/3	1/3	0.66
Alcool	1/3	4/3	9/3	4.67
Totaux	0.58	3.92	7.33	11.83

Total en colonne élevé
=
Groupe très différencié des autres

Total en ligne élevé
=
Critère permettant de différencier les groupes

AFC : importance du χ^2

- Le tableau des écarts pondérés est celui utilisé par l'AFC

	21-30 ans	31-40 ans	41-50 ans	Totaux
Goût	1/4	9/4	16/4	6.50
Couleur	0/3	1/3	1/3	0.66
Alcool	1/3	4/3	9/3	4.67
Totaux	0.58	3.92	7.33	11.83

- Les 3 lignes donnent un nuage de points
- Les 3 colonnes donnent un nuage de points
- On projette les 2 nuages sur un même plan avec une même origine

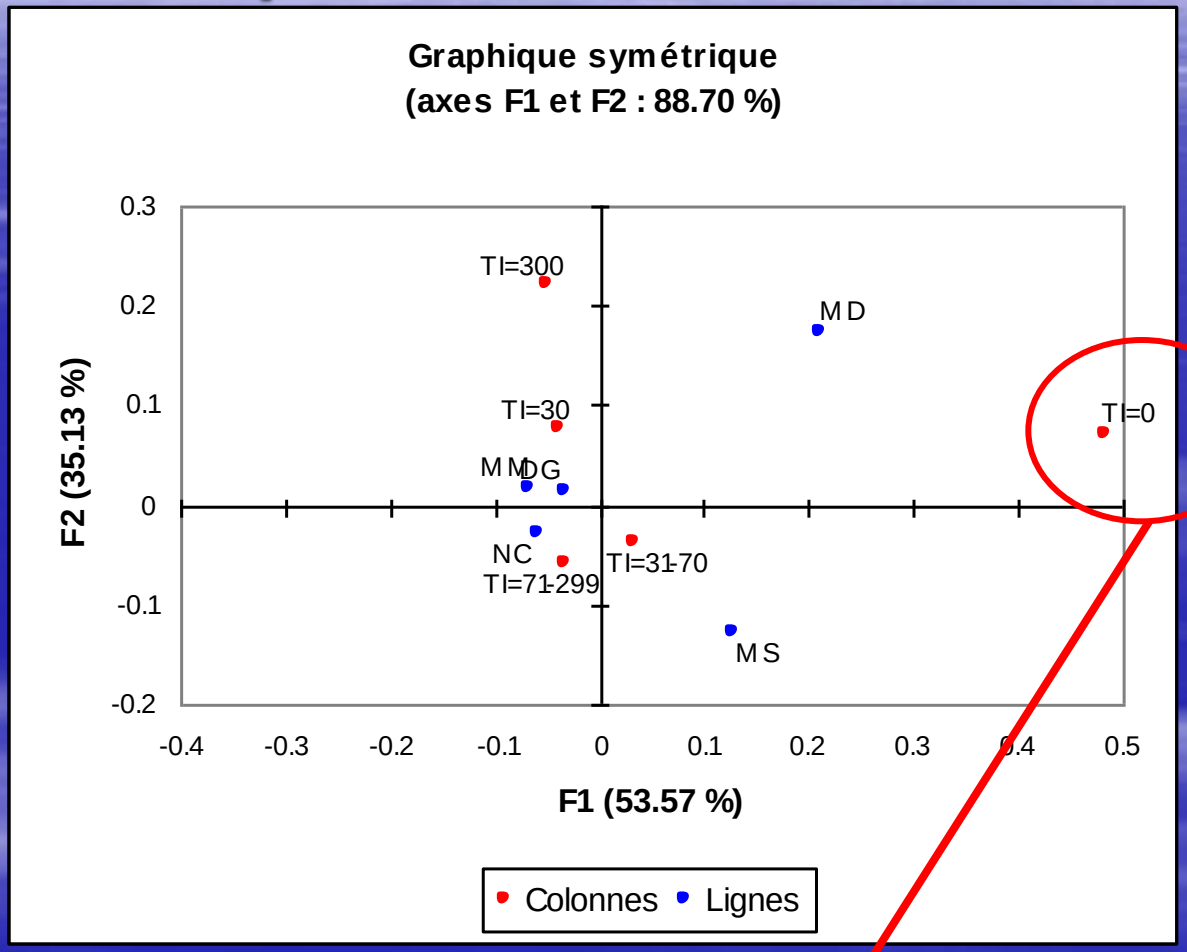
Ex : Comportement de la Caille

- Durée d'immobilité tonique découpée en 5 classes
- Cette variable est croisée avec l'identité du « preneur »

Preneur	TI=0	TI=1-30	TI=31-70	TI=71-299	TI=300	ENSEMBLE
DG	14	178	180	201	20	593
MD	6	30	29	24	4	93
MM	3	67	73	67	7	217
MS	8	44	67	65	3	187
NC	3	43	42	53	4	145
ENSEMBLE	34	362	391	410	38	1235

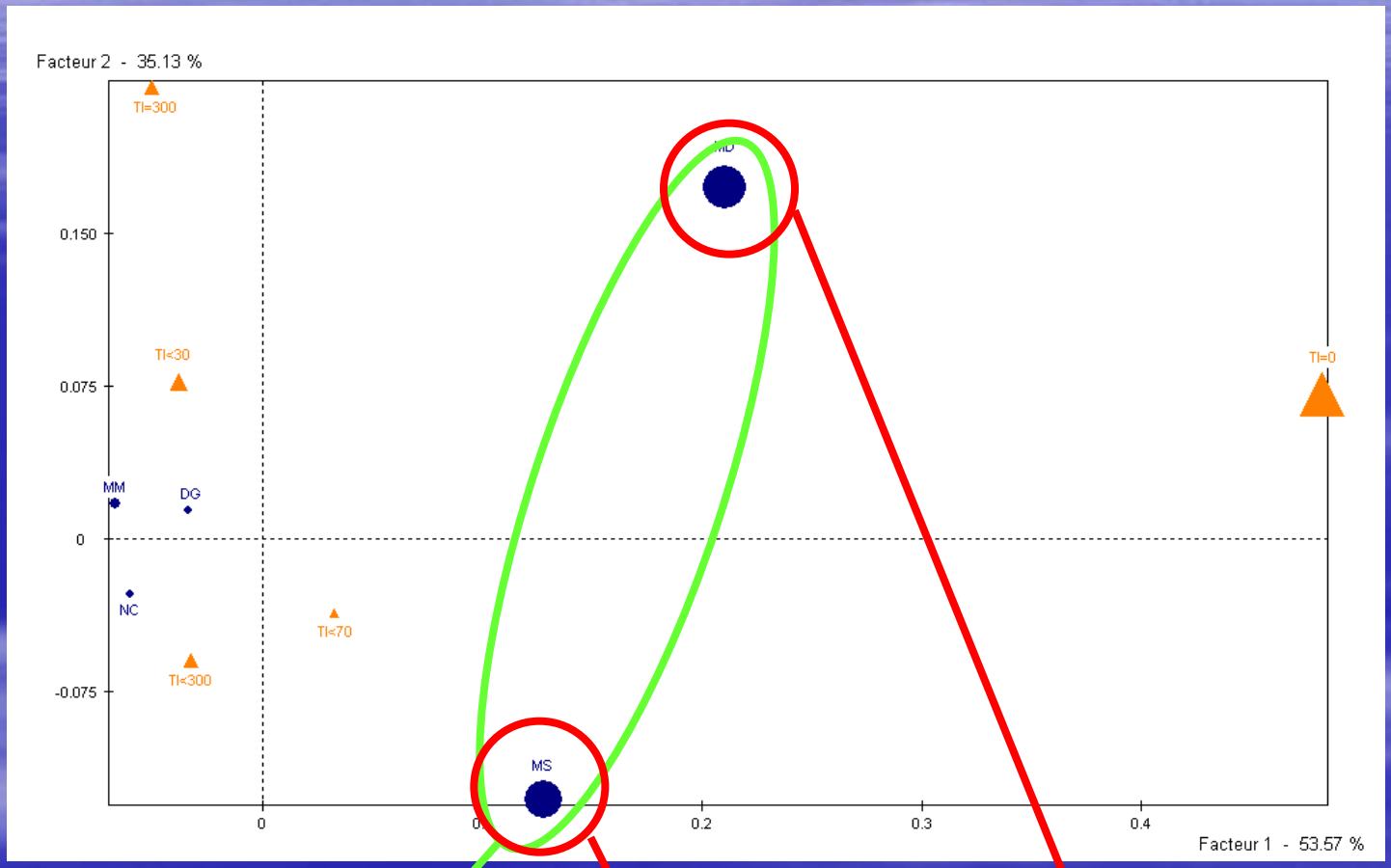
- Rentrer le tableau de contingence et réaliser l'AFC

Ex : Comportement de la Caille



F1 = opposition entre TI=0 et tout le reste

Ex : Comportement de la Caille



MS et MD ont relativement plus de TI=0 que les autres preneurs

Peu de TI=300

Beaucoup de TI=300

Opposition sur F2

Ex : Comportement de la Caille

Khi² (Valeur observée)	17.511
Khi² (Valeur critique)	26.296
DDL	16
p-value	0.353
alpha	0.05

Effectif global × Somme des valeurs propres

1235 0.015

suit un χ^2 à (nb lignes-1)(nb colonnes-1) d.d.l.

4 4

Interprétation du test :

H0 : Les lignes et les colonnes du tableau sont indépendantes.

Ha : Il existe un lien entre les lignes et les colonnes du tableau.

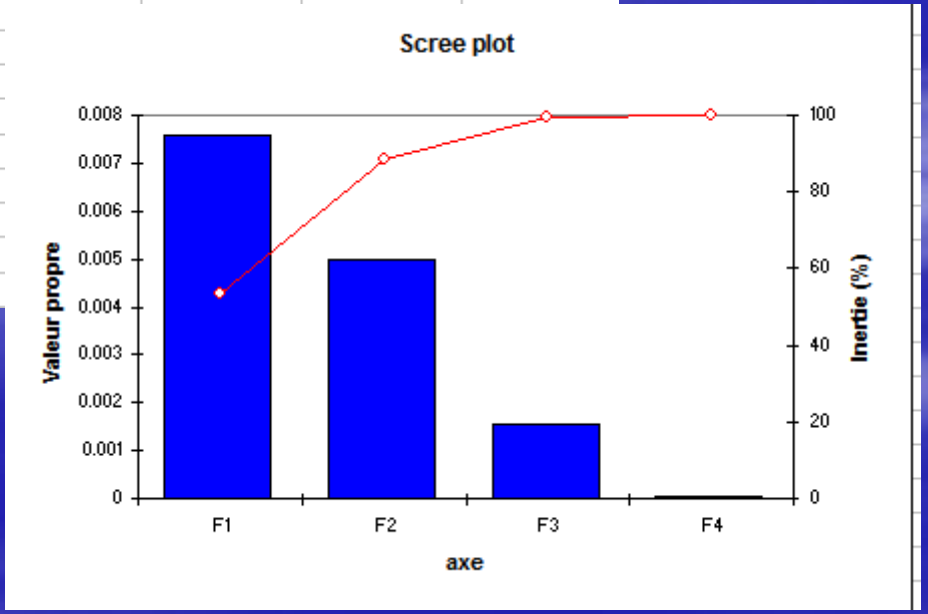
Etant donné que la p-value calculée est supérieure au niveau de signification seuil alpha=0.05, on peut valider l'hypothèse nulle H0.

Le risque de rejeter l'hypothèse nulle H0 alors qu'elle est vraie est de 35.33%.

Inertie totale : 0.014

Valeurs propres et pourcentages d'inertie :

	F1	F2	F3	F4
Valeur propre	0.008	0.005	0.002	0.000
Inertie (%)	53.569	35.132	10.952	0.347
% cumulé	53.569	88.701	99.653	100.000



Ex : Couleur des yeux et des cheveux

Couleur des yeux	Couleur des cheveux				
	Blond	Roux	Châtain	Marron	Noir
Pâles	326	38	241	110	3
Clairs	688	116	584	188	4
Moyens	343	84	909	412	26
Foncés	98	48	403	681	85

- Réaliser l'AFC
- Etape 1 : observer le χ^2

Ex : Couleur des yeux et des cheveux

Test d'indépendance entre les lignes et les colonnes :

Khi² (Valeur c	1240.039
Khi² (Valeur c	21.026
DDL	12
p-value	< 0.0001
alpha	0.05

Interprétation du test :

H0 : Les lignes et les colonnes du tableau sont indépendantes.

Ha : Il existe un lien entre les lignes et les colonnes du tableau.

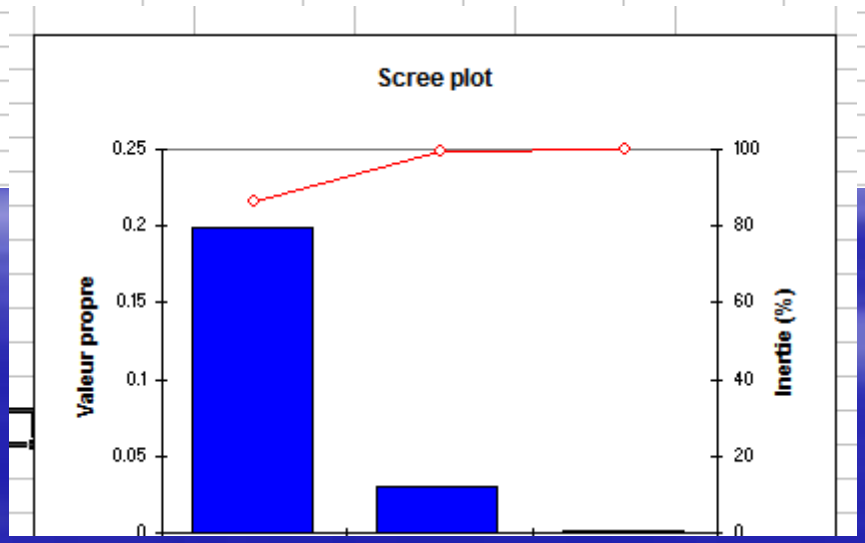
Etant donné que la p-value calculée est inférieure au niveau de signification alpha=0.05, on doit rejeter l'hypothèse nulle H0, et retenir l'hypothèse alternative Ha.

Le risque de rejeter l'hypothèse nulle H0 alors qu'elle est vraie est inférieur à 0.01%.

Inertie totale : 0.23

Valeurs propres et pourcentages d'inertie :

	F1	F2	F3
Valeur propre	0.199	0.030	0.001
Inertie (%)	86.556	13.070	0.373
% cumulé	86.556	99.627	100.000



Ex : Couleur des yeux et des cheveux

Résultats pour les yeux :

Poids, distances et distances quadratiques à l'origine, inerties et inerties relatives (yeux) :

	Poids (relatif)	Distance	Distance²	Inertie	Inertie relative
Pâles	0.133	0.438	0.192	0.02555	0.111
Clairs	0.293	0.451	0.203	0.05956	0.259
Moyens	0.329	0.247	0.061	0.02015	0.088
Foncés	0.244	0.715	0.512	0.12493	0.543

Coordonnées principales (yeux) :

	F1	F2	F3
Pâles	-0.400	0.165	-0.064
Clairs	-0.441	0.088	0.032
Moyens	0.034	-0.245	-0.006
Foncés	0.703	0.134	0.004

Coordonnées standard (yeux) :

	F1	F2	F3
Pâles	-0.897	0.954	-2.188
Clairs	-0.987	0.510	1.084
Moyens	0.075	-1.412	-0.189
Foncés	1.574	0.772	0.148

Contributions (yeux) :

	Poids (relatif)	F1	F2	F3
Pâles	0.133	0.107	0.121	0.638
Clairs	0.293	0.286	0.076	0.345
Moyens	0.329	0.002	0.657	0.012
Foncés	0.244	0.605	0.145	0.005

Cosinus carrés (yeux) :

	F1	F2	F3
Pâles	0.836	0.143	0.021
Clairs	0.956	0.039	0.005
Moyens	0.018	0.981	0.001
Foncés	0.965	0.035	0.000

Ex : Couleur des yeux et des cheveux

Résultats pour les cheveux :

Poids, distances et distances quadratiques à l'origine, inerties et inerties relatives (cheveux) :

	Poids (relatif)	Distance	Distance²	Inertie	Inertie relative
Blond	0.270	0.571	0.326	0.088	0.383
Roux	0.053	0.266	0.071	0.004	0.016
Châtain	0.397	0.213	0.045	0.018	0.078
Marron	0.258	0.598	0.357	0.092	0.401
Noir	0.022	1.132	1.282	0.028	0.122

Coordonnées principales (cheveux) :

	F1	F2	F3
Blond	-0.544	0.174	-0.013
Roux	-0.233	0.048	0.118
Châtain	-0.042	-0.208	-0.003
Marron	0.589	0.104	-0.010
Noir	1.094	0.286	0.046

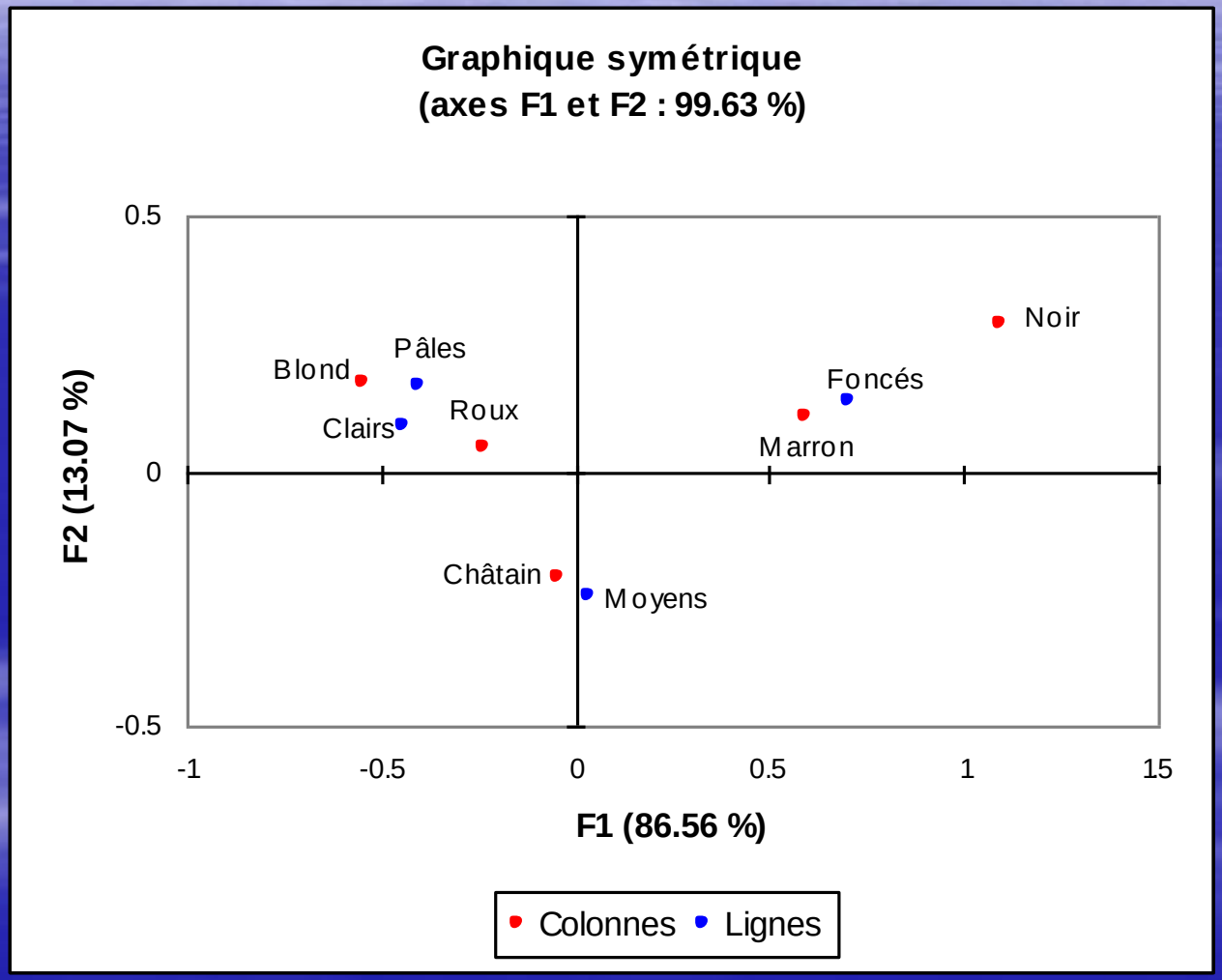
Coordonnées standard (cheveux) :

	F1	F2	F3
Blond	-1.219	1.002	-0.427
Roux	-0.523	0.278	4.027
Châtain	-0.094	-1.201	-0.110
Marron	1.319	0.599	-0.345
Noir	2.452	1.651	1.574

Contributions (cheveux) :

	Poids (relatif)	F1	F2	F3
Blond	0.270	0.401	0.271	0.049
Roux	0.053	0.014	0.004	0.861
Châtain	0.397	0.004	0.572	0.005
Marron	0.258	0.449	0.093	0.031
Noir	0.022	0.132	0.060	0.054

Ex : Couleur des yeux et des cheveux



ANALYSE DES CORRESPONDANCES MULTIPLES

- Même principe que l'AFC, mais on a plusieurs variables qualitatives pour chaque individu
- Le tableau contient alors les modalités des variables qualitatives, et non plus les fréquences de chaque modalité

Individu	Sexe	Lot	Age	…
1	M	1	<20	…
2	F	1	20-40	…
3	M	2	>60	…
4	M	2	40-60	…
5	F	3	<20	…
…	…	…	…	…

ANALYSE DES CORRESPONDANCES MULTIPLES

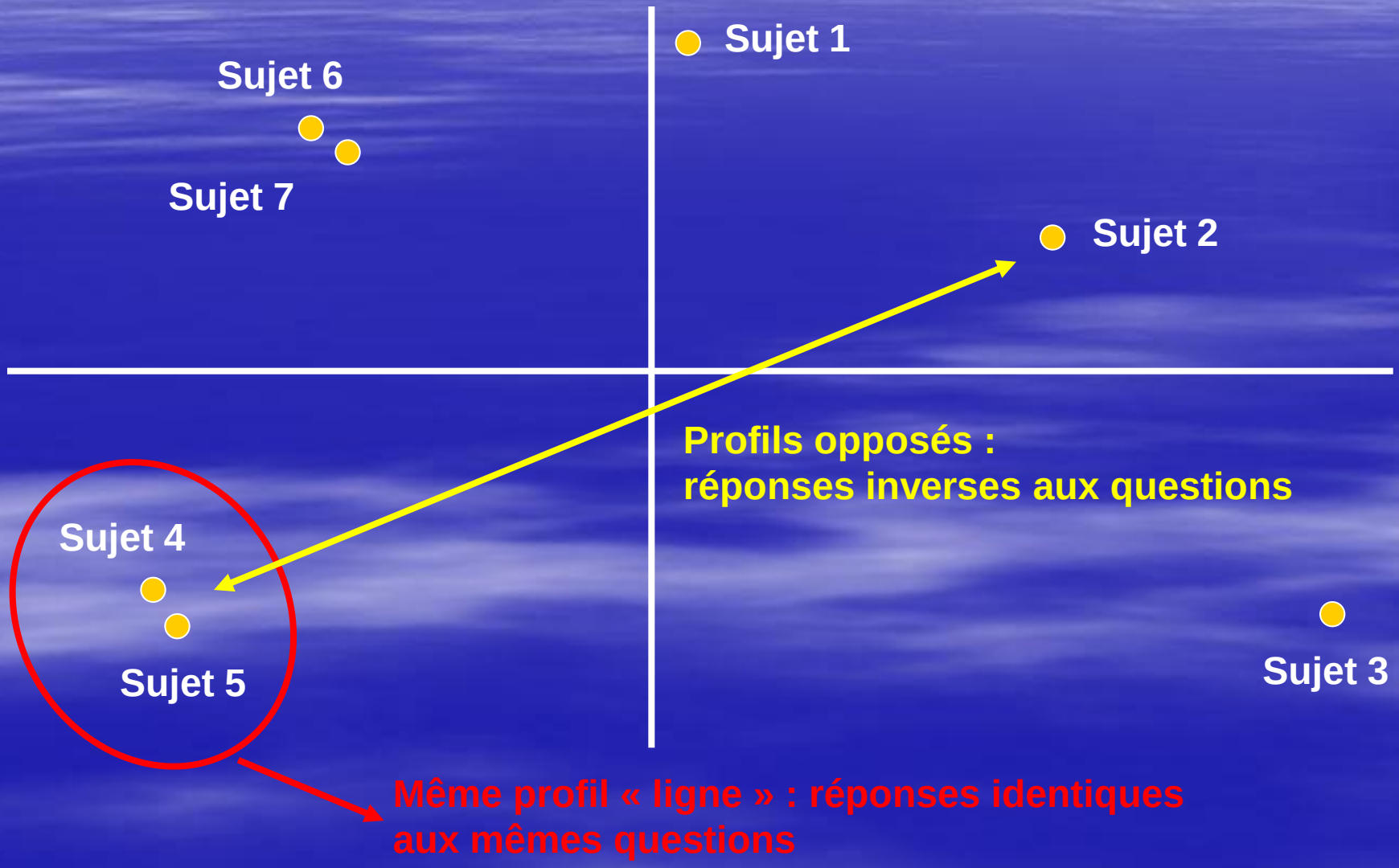
- On généralise les calculs de l'AFC
- A partir du tableau de départ, on crée des tableaux de « 0-1 » (tableaux disjonctifs complets)

Ind	Sexe	Lot	Age
1	M	1	<20
2	F	1	20-40
3	M	2	>60
4	M	2	40-60
5	F	3	<20

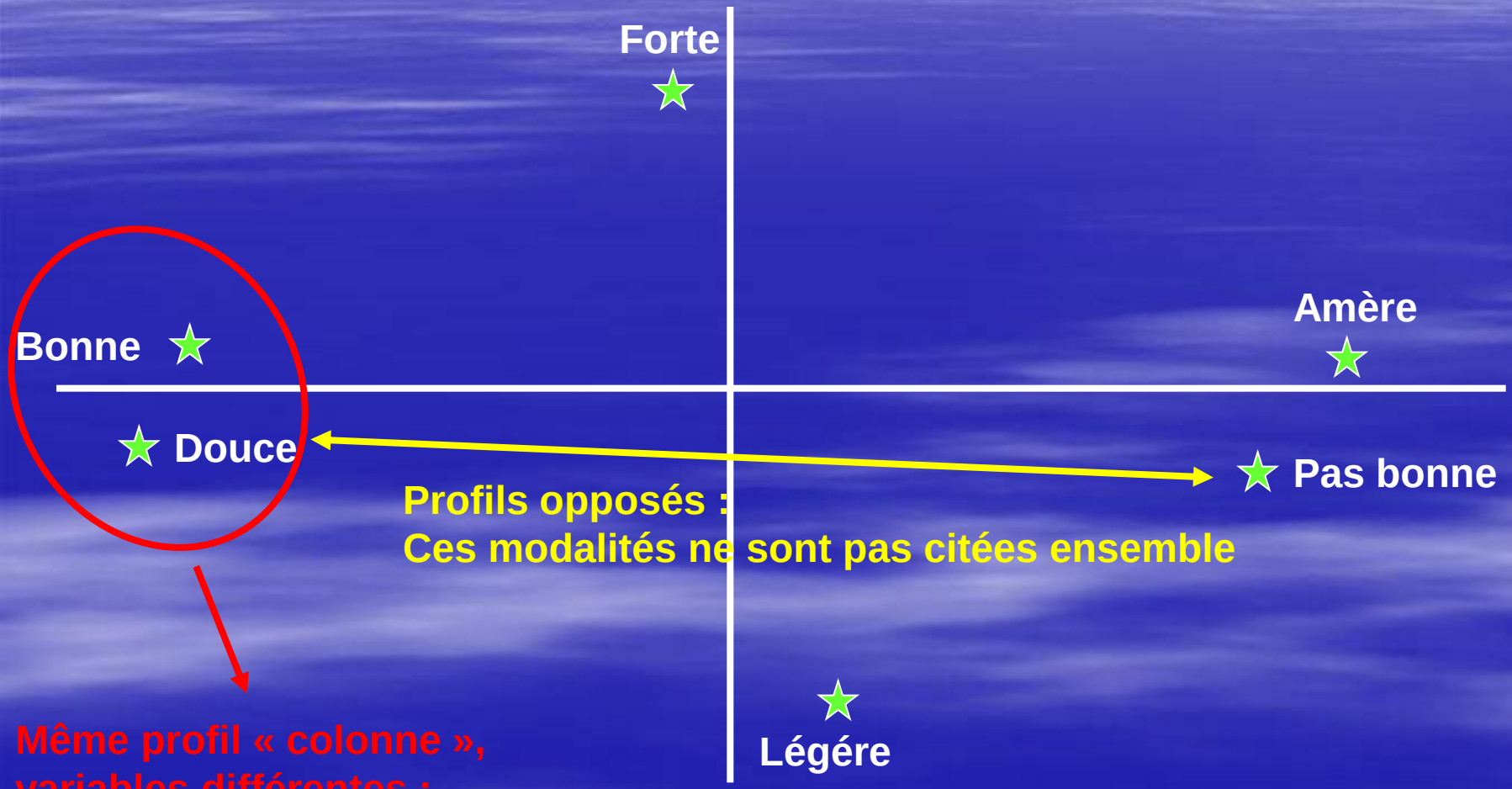
Mieux vaut avoir le moins de « 0 » possible, l'analyse est plus robuste

Ind	M	F	1	2	3	<20	20-40	40-60	>60
1	1	0	1	0	0	1	0	0	0
2	0	1	1	0	0	0	1	0	0
3	1	0	0	1	0	0	0	0	1
4	1	0	0	1	0	0	0	1	0
5	0	1	0	0	1	1	0	0	0

ANALYSE DES CORRESPONDANCES MULTIPLES : nuage d'individus



ANALYSE DES CORRESPONDANCES MULTIPLES : nuage de modalités



APPLICATION : CANCER DU SEIN

- Ouvrir la Base Sein. SBA
- Variables :
- Soins, indicateurs physiologiques :
  - chimio,lymphocytes, albumine, lactico dehydrogenase
- Indicateurs de gravité de la maladie :
  - qualitatives : délai d'apparition des métastases, nombre de sites avec métastase, métastases sur le foie
  - quantitatives : survie en mois

APPLICATION : CANCER DU SEIN

- Réaliser l'AFC
- Choisir la méthode :
  - Analyses Factorielles
  - Correspondances multiples (CORMU)
- Paramétrer l'AFC :
  - Nominales actives : chimio, lymphocytes, lactico dehydrogenase, albumine
  - Nominales illustratives : nb de sites métastatiques, délai d'apparition des métastases, métastases sur le foie
  - Continues illustratives : survie

APPLICATION : CANCER DU SEIN

VALEURS PROPRES

APERCU DE LA PRECISION DES CALCULS : TRACE AVANT DIAGONALISATION .. 1.2500

SOMME DES VALEURS PROPRES 1.2500

HISTOGRAMME DES 5 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	%	% CUMULE	
1	0.3174	25.39	25.39	**********
2	0.2768	22.14	47.54	*********
3	0.2439	19.51	67.05	********
4	0.2247	17.97	85.02	********
5	0.1872	14.98	100.00	*******

RECHERCHE DE PALIERS ENTRE (DIFFERENCES SECONDES)

PALIER ENTRE	VALEUR DU PALIER	
1 -- 2	7.74	**********

APPLICATION : CANCER DU SEIN

Tableau des valeurs propres

Trace de la matrice: 1.25000

Numéro	Valeur propre	Pourcentage	Pourcentage cumulé
1	0.3174	25.39	25.39
2	0.2768	22.14	47.54
3	0.2439	19.51	67.05
4	0.2247	17.97	85.02
5	0.1872	14.98	100.00

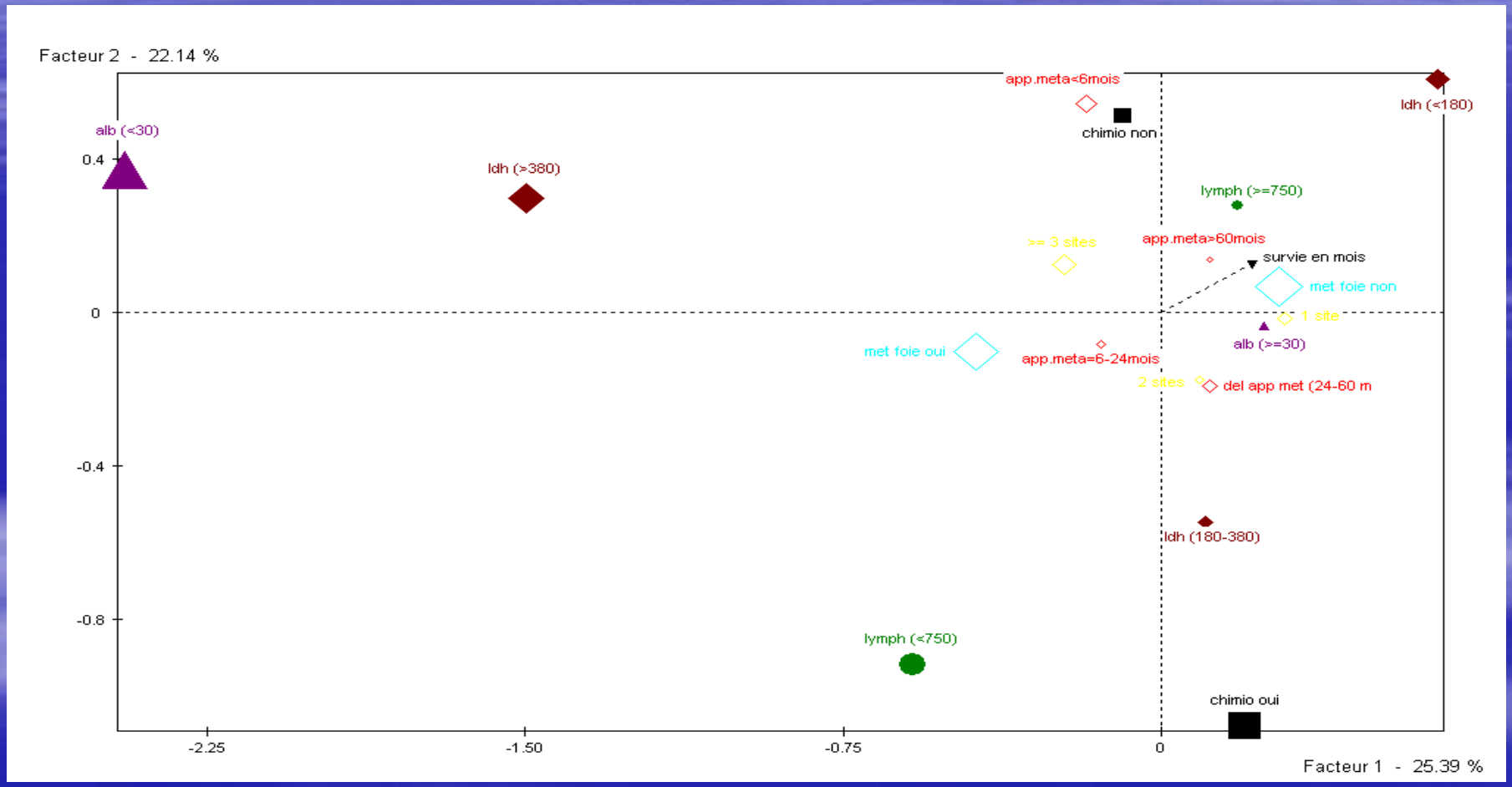
Contributions des modalités actives

Libellé	Poids relatif	Distance à l'origine	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5
chimio							
chimio non	16.978	0.47250	0.46	16.02	0.95	12.45	2.22
chimio oui	8.022	2.11640	0.98	33.89	2.01	26.34	4.69
lactico dehydrogenase							
ldh (<180)	8.574	1.91584	11.35	11.33	28.79	0.01	14.22
ldh (180-380)	11.885	1.10357	0.41	12.99	31.92	7.14	0.00
ldh (>380)	4.542	4.50467	32.05	1.45	3.12	17.55	27.65
lymphocytes							
lymph (>=750)	19.185	0.30310	1.91	5.37	7.28	8.49	0.21
lymph (<750)	5.815	3.29927	6.30	17.73	24.00	28.01	0.70
albumine							
alb (>=30)	22.750	0.09888	4.19	0.11	0.17	0.00	4.53
alb (<30)	2.250	10.11320	42.36	1.10	1.76	0.00	45.78

Valeurs-Tests des modalités actives et illustratives

Libellé	Effectif	Poids absolu	Distance à l'origine	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5
chimio								
chimio non	400	400.00	0.47250	-3.28	18.03	4.12	-14.32	5.51
chimio oui	189	189.00	2.11640	3.28	-18.03	-4.12	14.32	-5.51
lactico dehydrogenase								
ldh (<180)	202	202.00	1.91584	11.36	10.60	-15.86	0.28	-9.76
ldh (180-380)	280	280.00	1.10357	2.41	-12.70	18.68	-8.48	-0.14
ldh (>380)	107	107.00	4.50467	-17.10	3.40	-4.68	10.65	12.20
lymphocytes								
lymph (>=750)	452	452.00	0.30310	7.83	12.26	13.40	13.89	-2.00
lymph (<750)	137	137.00	3.29927	-7.83	-12.26	-13.40	-13.89	2.00
albumine								
alb (>=30)	536	536.00	0.09888	18.64	-2.81	-3.33	0.11	14.88
alb (<30)	53	53.00	10.11320	-18.64	2.81	3.33	-0.11	-14.88
delai apparition metastases								
< 6mois	84	84.00	6.01190	-1.74	5.37	0.42	-1.62	2.34
6-24 mois	164	164.00	2.59146	-2.18	-1.30	-2.22	1.65	0.80
24-60 mois	237	237.00	1.48523	2.24	-3.82	0.32	1.16	-2.35

APPLICATION : CANCER DU SEIN



Ex : Les voitures des Américains

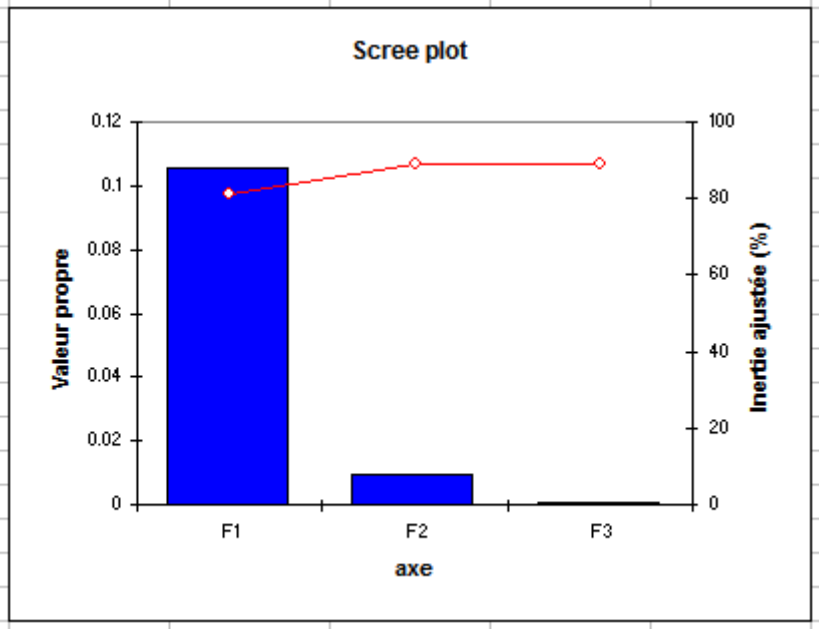
- Récupérer le fichier « voiture.xls »
- Réaliser les statistiques élémentaires sur les variables
- Réaliser l'analyse des correspondances multiples (ACM)

Ex : Les voitures des Américains

Inertie totale : 2

Valeurs propres et pourcentages d'inertie :

	F1	F2	F3	F4	F5	F6
Valeur propre	0.550	0.399	0.348	0.277	0.251	0.175
Inertie (%)	27.486	19.935	17.419	13.853	12.558	8.749
% cumulé	27.486	47.421	64.840	78.693	91.251	100.000
Inertie ajustée	0.105	0.010	0.001			
Inertie ajustée	81.366	7.427	0.393			
% cumulé	81.366	88.793	89.186			



Ex : Les voitures des Américains

Résultats pour les variables :

Coordonnées principales (Variables) :

	F1	F2	F3
Origine-Américaine	0.976	0.070	0.120
Origine-Européenne	-0.471	-1.048	1.907
Origine-Japonaise	-0.629	0.231	-0.613
Taille-Grande	1.738	0.957	1.006
Taille-Moyenne	0.278	-0.777	-0.421
Taille-Petite	-0.747	0.452	0.108
Type-Familiale	0.539	-0.299	-0.372
Type-Sportive	-0.878	-0.370	0.519
Type-Utilitaire	-0.025	1.713	0.192

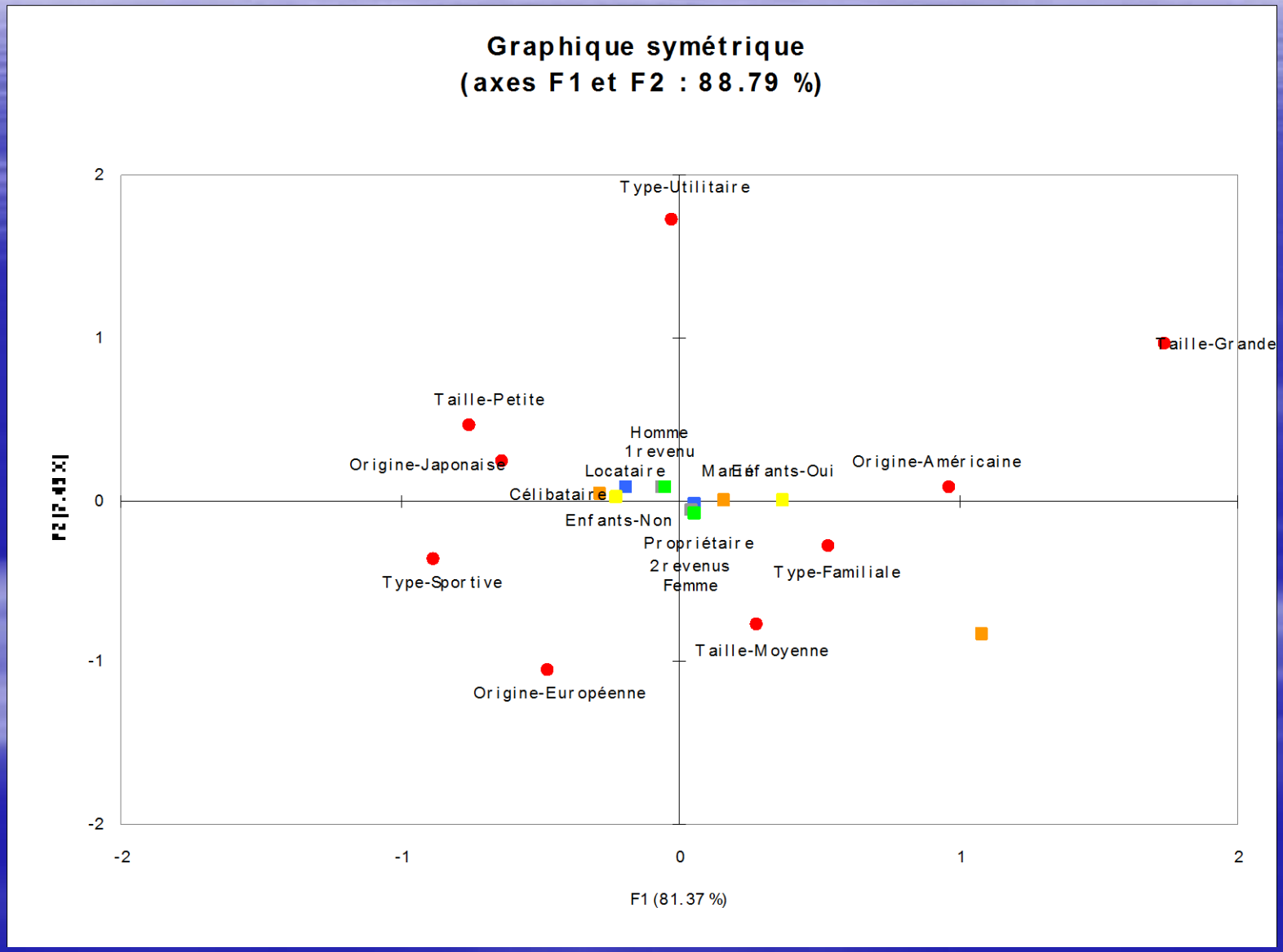
Contributions (Variables) :

	Poids	Poids (relatif)	F1	F2	F3
Origine-Américaine	128	0.126	0.219	0.002	0.005
Origine-Européenne	45	0.044	0.018	0.122	0.463
Origine-Japonaise	165	0.163	0.117	0.022	0.176
Taille-Grande	43	0.042	0.233	0.097	0.123
Taille-Moyenne	142	0.140	0.020	0.212	0.071
Taille-Petite	153	0.151	0.153	0.077	0.005
Type-Familiale	177	0.175	0.092	0.039	0.069
Type-Sportive	107	0.106	0.148	0.036	0.081
Type-Utilitaire	54	0.053	0.000	0.392	0.006

Ex : Les voitures des Américains

Résultats pour les variables :			
Coordonnées principales (Variables) :			
Domicile-.	0.889	-0.035	0.264
Domicile-Locataire	-0.187	0.061	-0.165
Domicile-Propriétaire	0.060	-0.023	0.059
Revenu-1 revenu	-0.057	0.077	0.049
Revenu-2 revenus	0.045	-0.061	-0.039
Statut marital-	-0.275	0.029	0.072
Statut marital-Marié	0.162	-0.017	-0.042
Enfants-.	1.087	-0.841	-0.644
Enfants-Non	-0.227	0.013	0.069
Enfants-Oui	0.371	-0.015	-0.110
Sexe-Femme	0.061	-0.090	-0.023
Sexe-Homme	-0.050	0.074	0.019

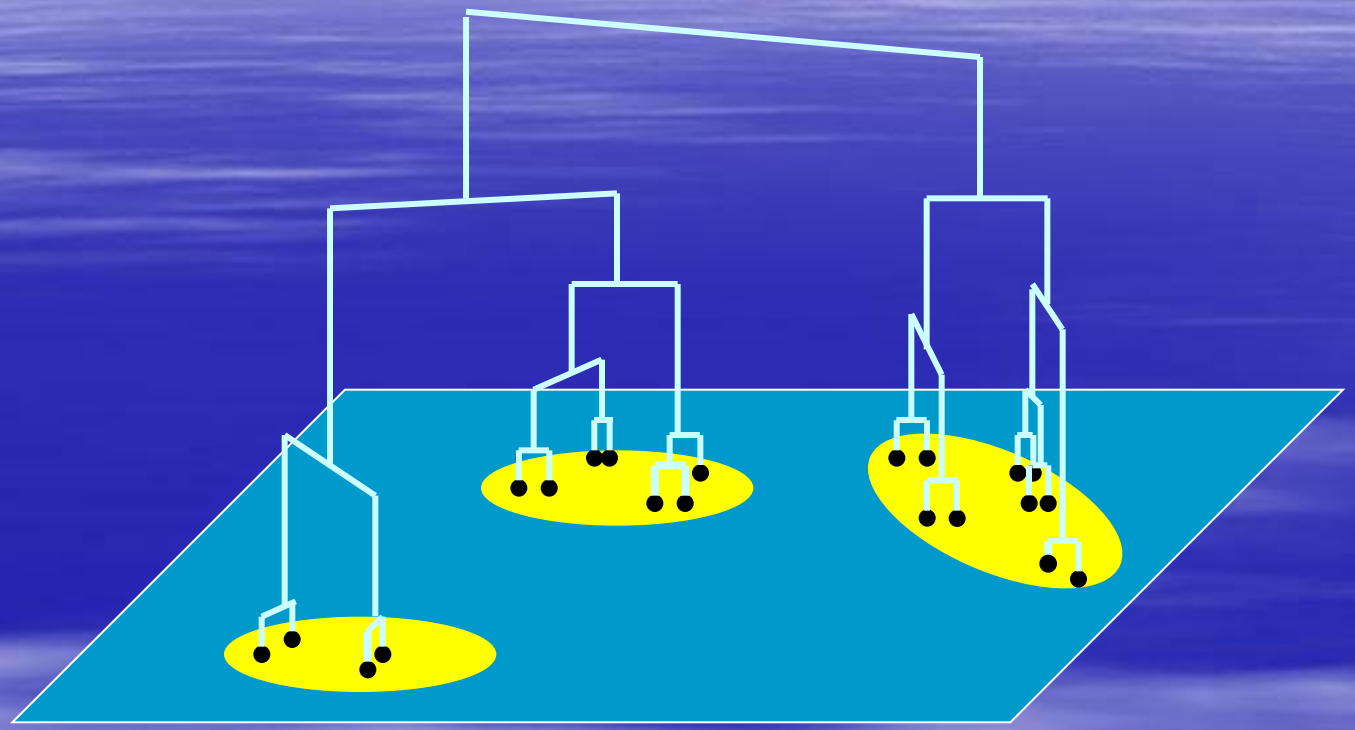
Ex : Les voitures des Américains



Classification ascendante hiérarchique

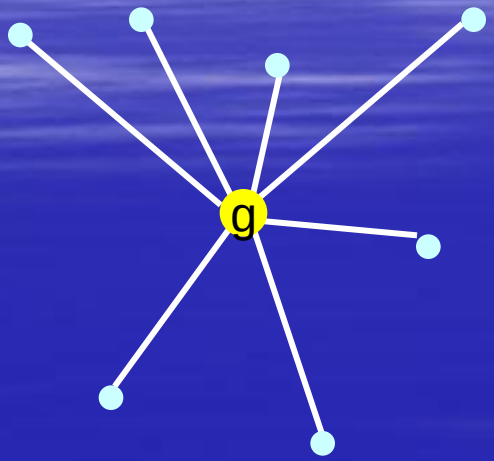
- Question : peut-on faire des groupes à partir de nos données, pour prévoir où un nouvel individu ira se placer ?

Principe



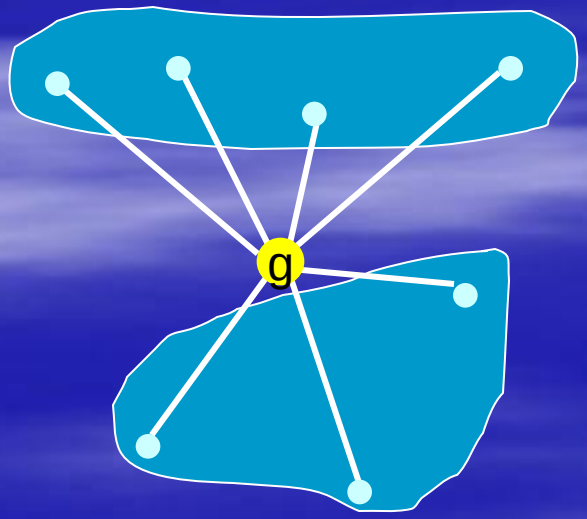
On recherche les deux points les plus proches
On les regroupe, et leur barycentre devient un nouveau point
On continue jusqu'à n'avoir plus qu'une seule classe

Critère de Ward



Inertie totale =

Distance au centre de gravité



Inertie inter classes + Inertie intra-classe

Agrégation : les variables

- Sur les variables quantitatives
- Sur les données brutes
  - lourdeur si beaucoup d'individus
  - données qualitatives
- Utiliser les coordonnées factorielles
- Sur données qualitatives, réaliser une AFC d'abord, puis utiliser les coordonnées comme base d'agrégation