ACSEG 2004
11ème Rencontre Internationale
Lille, 18 & 19 novembre 2004

Traitements de données qualitatives par des algorithmes fondés sur l'algorithme de Kohonen

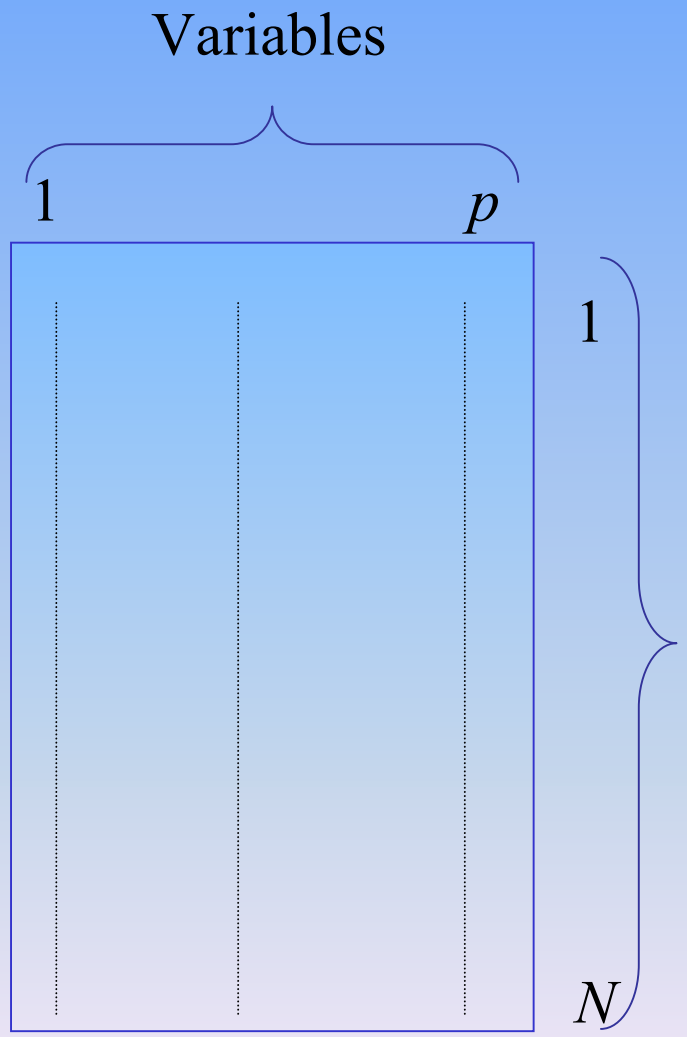
Patrick Letrémy

SAMOS-MATISSE

CNRS UMR 8595

Université Paris 1- Sorbonne

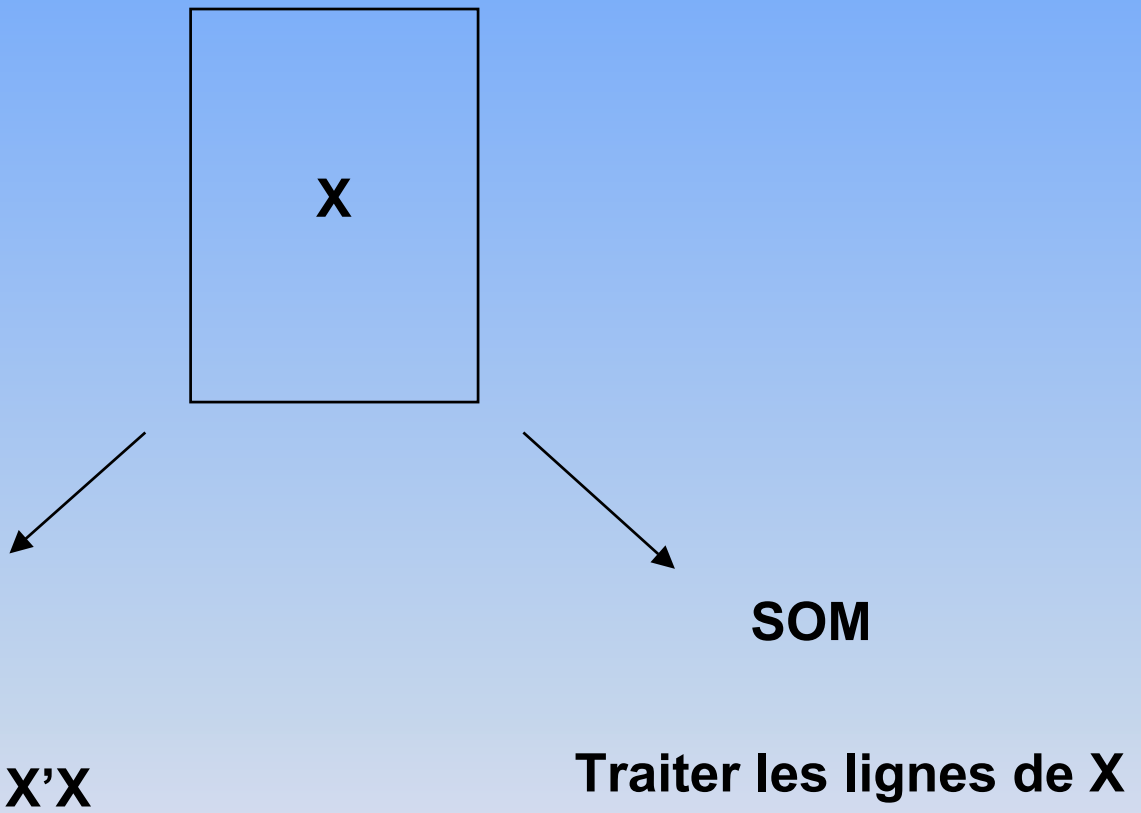
Analyse de données



Comment extraire
de l'information ?

Si N et p sont grands
Réduire p : *Analyse Factorielle*
Réduire N : *Classification*

Parallèle entre Carte de Kohonen (SOM) et ACP



Croisement d'une classification avec une variable qualitative : exemple de 96 pays en 1996

Les 7 variables quantitatives

- ANCRX Croissance annuelle de la population en %
- TXMORT Taux de mortalité infantile (en pour mille)
- TXANAL Taux d'analphabétisme en %
- SCOL2 Indice de fréquentation scolaire au second degré
- PNBH PNB par habitant exprimé en dollars
- CHOMAG Taux de chomage en %
- INFLAT Taux d'inflation en %

La variable qualitative

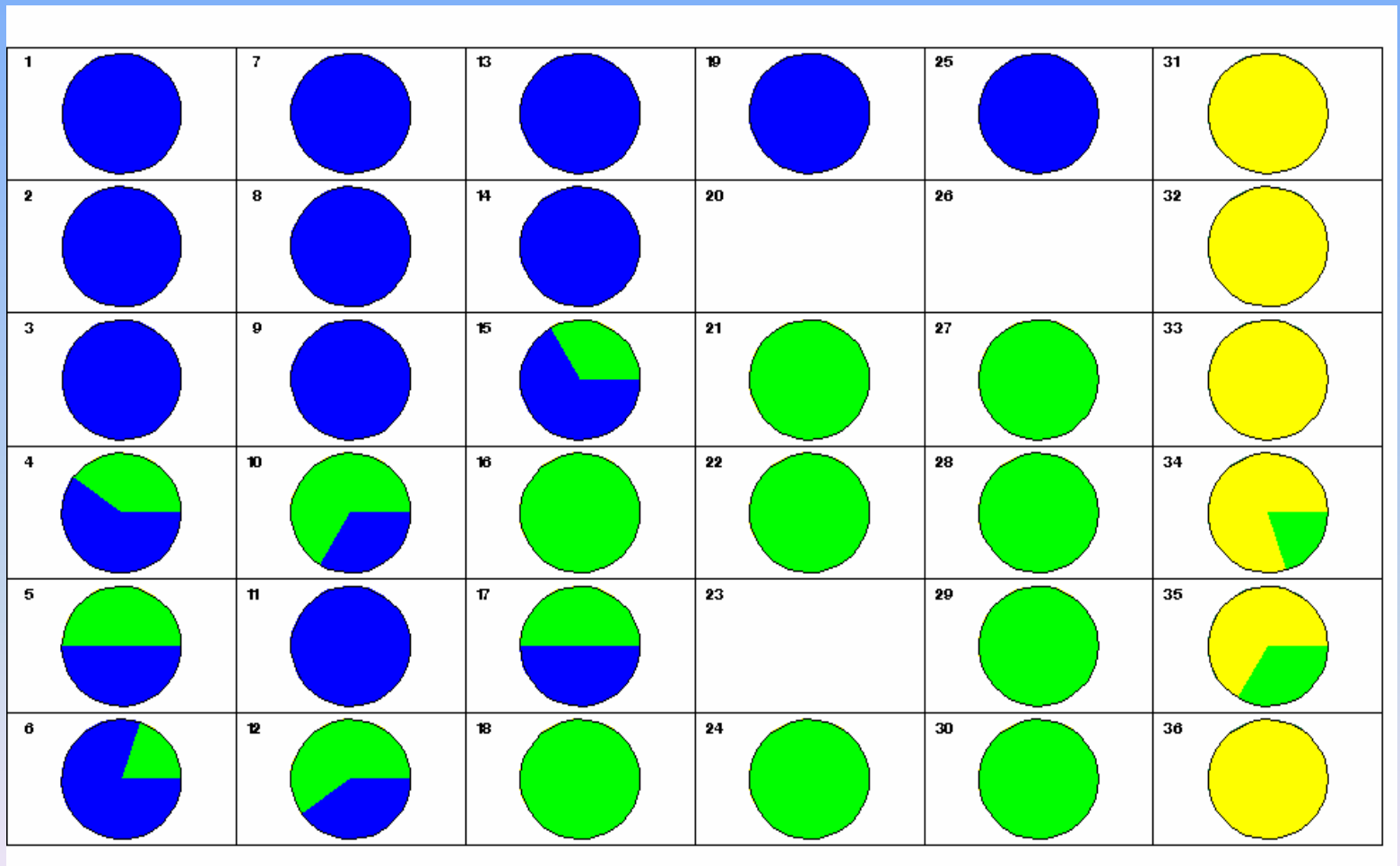
- IDH Indice du Développement Humain (3 niveaux: faible, moyen, élevé)

Les 36 classes de Kohonen, regroupées en 7 macro-classes, 600 itérations

Japan Switzerland	USA Luxembourg	Cyprus South Korea	Russia Ukraine	Brazil	Angola
Germany Belgium Denmark France Norway Netherlands Sweden	Australia Canada Iceland	U Arab Emirates Israel Singapore			Afghanistan Mozambique Yemen
Spain Finland Ireland Italy New-Zealand United Kingdom	Greece Malta Portugal	Argentina Bahrain Philippines	Mongolia Peru	Swaziland	Mauritania Sudan
Croatia Hungary Romania Slovenia R. Czech	Bulgaria Poland Yugoslavia	Jamaica Sri Lanka	Tunisia Turkey	Saudi Arabia Bolivia El Salvador Syria	Comoros Ivory Coast Ghana Morocco Pakistan
Moldavia Uruguay	Chili Colombia	Albania Panama		Algeria Egypt Nicaragua	Cameroon Laos Nigeria
China Fiji Malaysia Mexico Thailand	Costa Rica Ecuador Guyana Paraguay Venezuela	Vietnam	South Africa Lebanon Macedonia	Indonesia Iran Namibia Zimbabwe	Haiti Kenya

Répartition de la variable IDH sous forme sectorielle pour chaque cellule de la grille 6x6

jaune (faible), vert (moyen), bleu (élevé)



Enquête constituée uniquement de variables qualitatives

Enquête de *N* individus et *K* variables qualitatives (questions).

Chaque question *k* ($1 \leq k \leq K$) possède m_k modalités.

Les individus répondent à chaque question *k* en choisissant seulement une modalité parmi les m_k modalités.

M est le nombre total de modalités

Si on veut savoir qui répond quoi, on utilise le *Tableau Disjonctif Complet*, matrice D de format NxM.

Si on veut seulement étudier les relations entre les K variables, on utilise la *table de Burt*, matrice symétrique B de format MxM définie par B=D'D.

Exemple : enquête de N=20 individus, K=3 questions Question1 (3 modalités), Question2 (2), Question3 (3) au total M=8 modalités

Enquete	Question1	Question2	Question3
indiv1	2	2	1
indiv2	1	1	2
indiv3	1	2	1
indiv4	2	1	3
indiv5	3	1	3
indiv6	1	1	3
indiv7	1	2	1
indiv8	2	2	2
indiv9	2	2	2
indiv10	3	1	1
indiv11	3	1	1
indiv12	3	2	3
indiv13	1	1	2
indiv14	2	1	1
indiv15	1	1	2
indiv16	1	1	2
indiv17	1	1	2
indiv18	3	1	3
indiv19	3	1	3
indiv20	3	1	3

Tableau Disjonctif Complet de format (20x8)

	Question1			Question2		Question3		
Indiv	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**
indiv1	0	1	0	0	1	1	0	0
indiv2	1	0	0	1	0	0	1	0
indiv3	1	0	0	0	1	1	0	0
indiv4	0	1	0	1	0	0	0	1
indiv5	0	0	1	1	0	0	0	1
indiv6	1	0	0	1	0	0	0	1
indiv7	1	0	0	0	1	1	0	0
indiv8	0	1	0	0	1	0	1	0
indiv9	0	1	0	0	1	0	1	0
indiv10	0	0	1	1	0	1	0	0
indiv11	0	0	1	1	0	1	0	0
indiv12	0	0	1	0	1	0	0	1
indiv13	1	0	0	1	0	0	1	0
indiv14	0	1	0	1	0	1	0	0
indiv15	1	0	0	1	0	0	1	0
indiv16	1	0	0	1	0	0	1	0
indiv17	1	0	0	1	0	0	1	0
indiv18	0	0	1	1	0	0	0	1
indiv19	0	0	1	1	0	0	0	1
indiv20	0	0	1	1	0	0	0	1

Table de BURT de format 8x8

		Question1			Question2		Question3		
		1	2	3	1	2	1	2	3
Question1	1	8	0	0	6	2	2	5	1
	2	0	5	0	2	3	2	2	1
	3	0	0	7	6	1	2	0	5
Question2	1	6	2	6	14	0	3	5	6
	2	2	3	1	0	6	3	2	1
Question3	1	2	2	2	3	3	6	0	0
	2	5	2	0	5	2	0	7	0
	3	1	1	5	6	1	0	0	7

Analyse Factorielle des Correspondances (AFC) et Analyse des Correspondances Multiples (ACM)

L' ACM est une généralisation de l' AFC.

L' AFC traite un *tableau de contingence*.

Soient *deux variables qualitatives* ayant respectivement *I* et *J* modalités.

Le tableau de contingence pour ces deux variables est une matrice *IxJ*, où l'entrée n_{ij} est le nombre d'individus qui partagent la modalité *i* pour la première variable (variable ligne) et la modalité *j* pour la seconde (variable colonne).

Ce cas est fondamental, puisqu'à la fois *le tableau Disjonctif Complet D* et *l table de Burt B* peuvent être considérés comme des tableaux de contingence.

D est le tableau de contingence qui croise la « méta-variable » INDIVIDU à *N* valeurs avec la « méta-variable » MODALITE de *M* valeurs.

B est le tableau de contingence qui croise la « méta-variable » MODALITE de *M* valeurs avec elle-même.

Analyse Factorielle des Correspondances, appliquée au tableau de contingence.

- la table F des fréquences relatives, aux entrées $f_{ij} = \frac{n_{ij}}{n}$, avec $n = \sum_{i,j} n_{ij}$
- les marges aux entrées $f_{i\bullet} = \sum_j f_{ij}$ et $f_{\bullet j} = \sum_i f_{ij}$
- la table P_R des I profils lignes de somme 1, aux entrées $p_{ij}^R = \frac{f_{ij}}{f_{i\bullet}}$
- la table P_C des J profils colonnes de somme 1, aux entrées $p_{ij}^C = \frac{f_{ij}}{f_{\bullet j}}$

Ces profils forment deux ensembles de points respectivement dans R^J et R^I. Les moyennes de ces deux ensembles sont respectivement notées

$$\bar{i} = (f_{\bullet 1}, f_{\bullet 2}, \cdots, f_{\bullet J}) \quad \text{et} \quad \bar{j} = (f_{1\bullet}, f_{2\bullet}, \cdots, f_{I\bullet})$$

Ces profils sont en fait des distributions de probabilités conditionnelles, il est d'usage de leur appliquer la distance du chi-deux, définie comme suit
pour les lignes :

$$\chi^2(i,i') = \sum_j \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{i'j}}{f_{i'\bullet}} \right)^2 = \sum_j \left(\frac{f_{ij}}{\sqrt{f_{\bullet j}} f_{i\bullet}} - \frac{f_{i'j}}{\sqrt{f_{\bullet j}} f_{i'\bullet}} \right)^2$$

pour les colonnes :

$$\chi^2(j,j') = \sum_i \frac{1}{f_{i\bullet}} \left(\frac{f_{ij}}{f_{\bullet j}} - \frac{f_{ij'}}{f_{\bullet j'}} \right)^2 = \sum_j \left(\frac{f_{ij}}{\sqrt{f_{i\bullet}} f_{\bullet j}} - \frac{f_{ij'}}{\sqrt{f_{i\bullet}} f_{\bullet j'}} \right)^2$$

Notons que chaque ligne i est pondérée par $f_{i\bullet}$ et que chaque colonne j est pondérée par $f_{\bullet j}$.

AFC appliquée au tableau de contingence (suite)

Calcul de l'inertie de ces deux ensembles de profils

Pour les profils lignes $= \sum_i f_{i\bullet} \chi^2(i,\bar{i})$, pour les profils colonnes $= \sum_j f_{\bullet j} \chi^2(j,\bar{j})$

Elles sont égales. Cette inertie totale sera notée $\Im = \sum_{i,j} \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{f_{i\bullet} f_{\bullet j}} = \sum_{i,j} \frac{f_{ij}^2}{f_{i\bullet} f_{\bullet j}} - 1$

On peut souligner deux points importants :

1) Pour utiliser la distance euclidienne entre les lignes et entre les colonnes au lieu de la distance du khi-deux, et prendre en compte le poids de chaque ligne par $f_{i\bullet}$ et le poids de chaque colonne par $f_{\bullet j}$, on remplace les valeurs initiales f_{ij}

par les valeurs corrigées f_{ij}^C, en posant $\boxed{f_{ij}^C = \frac{f_{ij}}{\sqrt{f_{i\bullet}}\sqrt{f_{\bullet j}}}}$

On note F^C la matrice dont les entrées sont f_{ij}^C.

2) L'inertie totale est $\Im = \frac{1}{N} T$, où T est la statistique du chi-deux qui permet de tester l'indépendance entre la variable ligne et la variable colonne ; la statistique T mesure l'écart à l'indépendance.

AFC appliquée au tableau de contingence (fin)

L'Analyse Factorielle des Correspondances (AFC) est simplement une double ACP sur les lignes et les colonnes de la matrice corrigée F^C.

Pour les profils lignes, les valeurs propres et les vecteurs propres sont obtenus par diagonalisation de la matrice $F^c{}'F^c$.

Pour les profils colonnes, les valeurs propres et les vecteurs propres sont obtenus par diagonalisation de la matrice transposée $F^cF^c{}'$.

On sait que les deux matrices ont les mêmes valeurs propres et que leurs vecteurs propres sont liés. Il est facile de prouver que l'inertie totale $\Im$ est égale à la somme des valeurs propres de $F^c{}'F^c$ ou de $F^cF^c{}'$.

Ainsi l'AFC décompose l'écart à l'indépendance en une somme de termes décroissants associés aux axes principaux des deux ACP.

Pour l'AFC qui ne traite que deux variables, le couplage des deux ACP est garanti, car on utilise deux matrices transposées.

Il est ainsi possible de représenter simultanément les modalités des deux variables.

En Conclusion de L'AFC

Si on reprend le parallèle entre SOM et ACP.

La diagonalisation de la matrice $F^{c\prime}F^{c}$ peut être remplacée par l'algorithme SOM appliqué aux profils lignes.

La diagonalisation de $F^{c}F^{c\prime}$ peut être remplacée par l'algorithme SOM appliqué aux profils colonnes.

***C'est le point clé pour définir l'algorithme SOM adapté aux variables qualitatives*.**

L'Analyse des Correspondances Multiples (ACM)

1) *On s'intéresse uniquement aux modalités*.
Comme la table de Burt est un tableau de contingence, il suffit de faire une AFC sur *la table de Burt corrigée* B^c , avec

$$b_{jl}^c = \frac{b_{jl}}{\sqrt{b_{j\bullet}}\sqrt{b_{\bullet l}}} = \frac{b_{jl}}{K\sqrt{b_j}\sqrt{b_l}} \quad \text{puisque } b_{j\bullet} = Kb_j \text{ et } b_{l\bullet} = Kb_l$$

Comme les matrices B et B^c sont symétriques, les diagonalisations de $B^{c\prime}B^c$ ou $B^cB^{c\prime}$ sont identiques, et donnent une représentation simultanée des M modalités des K variables (questions).

2) *On s'intéresse aussi aux individus.*
Comme le tableau Disjonctif Complet est aussi un tableau de contingence, il suffit de faire une AFC sur *le tableau Disjonctif Complet corrigé* D^c , avec

$$d_{ij}^c = \frac{d_{ij}}{\sqrt{d_{i\bullet}}\sqrt{d_{\bullet j}}} = \frac{d_{ij}}{\sqrt{K}\sqrt{b_j}}$$

Dans ce cas, la matrice D^c n'est plus symétrique. La diagonalisation de $D^{c\prime}D^c$ donne une représentation des individus, celle de $D^cD^{c\prime}$ fournit une représentation des modalités.

Passage de l'ACM à l'algorithme SOM

Si on reprend la conclusion de L' AFC, il est très facile de définir de nouveaux algorithmes fondés sur l'algorithme SOM.

1) *Si l'on veut traiter uniquement des modalités*, comme la matrice de Burt est symétrique, il suffit d'utiliser SOM sur les lignes (ou les colonnes) de B^c pour obtenir une bonne représentation de toutes les modalités sur une carte de Kohonen. Cette remarque fonde la définition de l'algorithme KACM (Kohonen Analyse des Correspondances Multiples).

2) *Si l'on veut garder les individus*, on peut appliquer SOM aux lignes de D^c, mais on obtiendra une carte de Kohonen pour les seuls individus. Pour représenter simultanément les modalités, il est nécessaire de trouver une autre astuce.

Deux techniques sont définies :

a) KACM_ind (Kohonen Analyse des Correspondances Multiples avec individus) : les modalités sont affectées aux classes après apprentissage, comme données supplémentaires.

b) KDISJ (Kohonen sur tableau DISJonctif) : deux algorithmes SOM sont utilisés sur les lignes (individus) et sur les colonnes (modalités) de D^c, l'association entres modalités et individus est contrainte durant tout l'apprentissage.

Traitements des Modalités pour les 96 pays

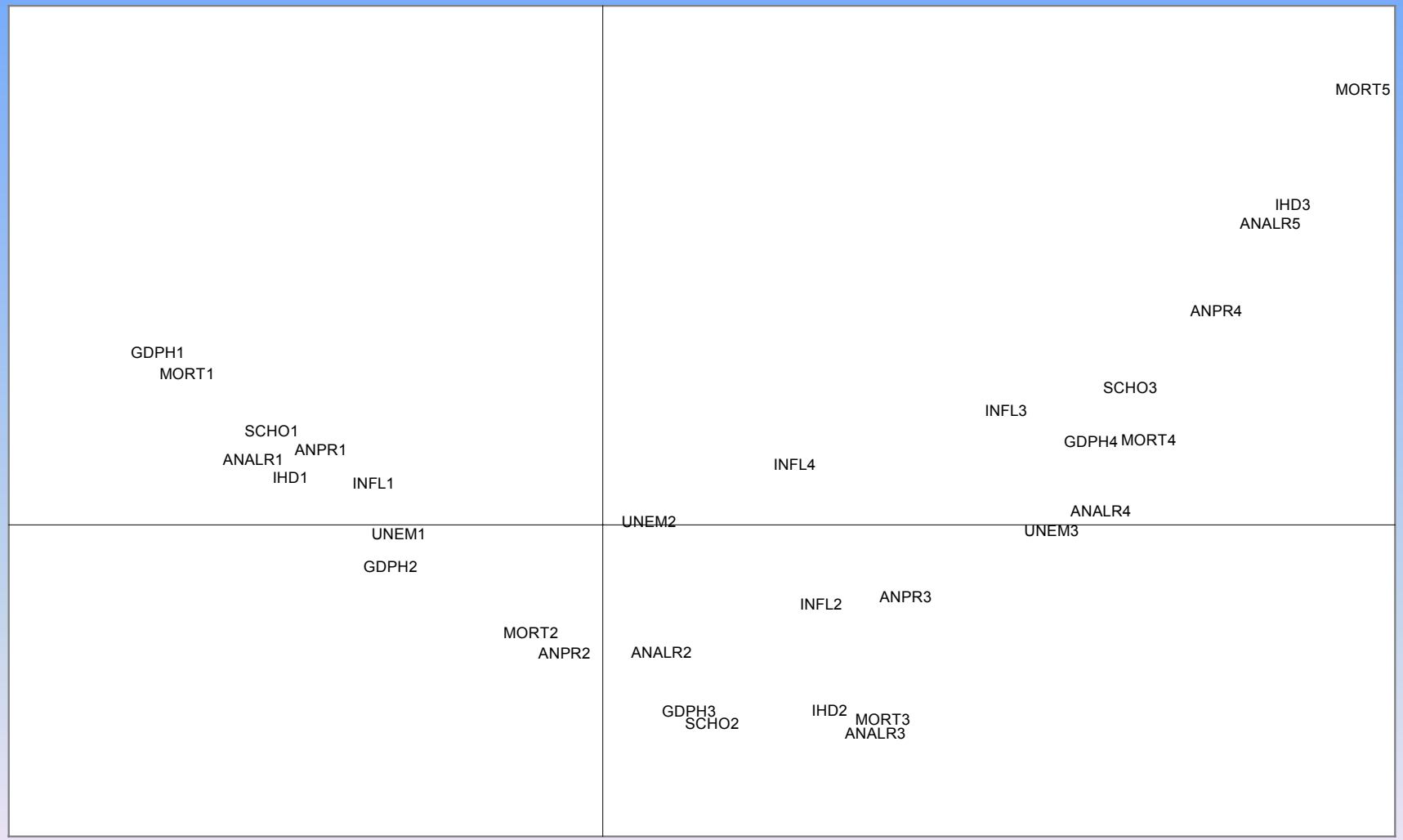
- On discrétise les 7 variables quantitatives en classes
- On ajoute la variable qualitative IDH
- Au total on obtient 8 variables qualitatives (questions) et 31 modalités

Variables	Tranches	Noms des modalités
Annual population growth	[-1, 1[, [1, 2[, [2, 3[, ≥3	ANPR1, ANPR2, ANPR3, ANPR4
Mortality rate	[4, 10[10, 40[40, 70[, [70, 100[,≥ 100	MORT1, MORT2, MORT3, MORT4, MORT5
Analphabetism rate	[0, 6[, [6, 20[, [20, 35[, [35, 50[, ≥ 50	ANALR1, ANALR2, ANALR3, ANALR4, ANALR5
High school	≥ 80, [40, 80[, [4, 40[	SCHO1, SCHO2, SCHO3
GDPH	≥ 10000, [3000,10000[, [1000, 3000[, < 1000	GDPH1, GDPH2, GDPH3, GDPH4
Unemployment rate	[0, 10[, [10, 20[, ≥ 20	UNEM1, UNEM2, UNEM3
Inflation rate	[0, 10[, [10, 50[, [50, 100[, ≥ 100	INFL1, INFL2, INFL3, INFL4
IHD	1, 2, 3	IHD1, IHD2, IHD3

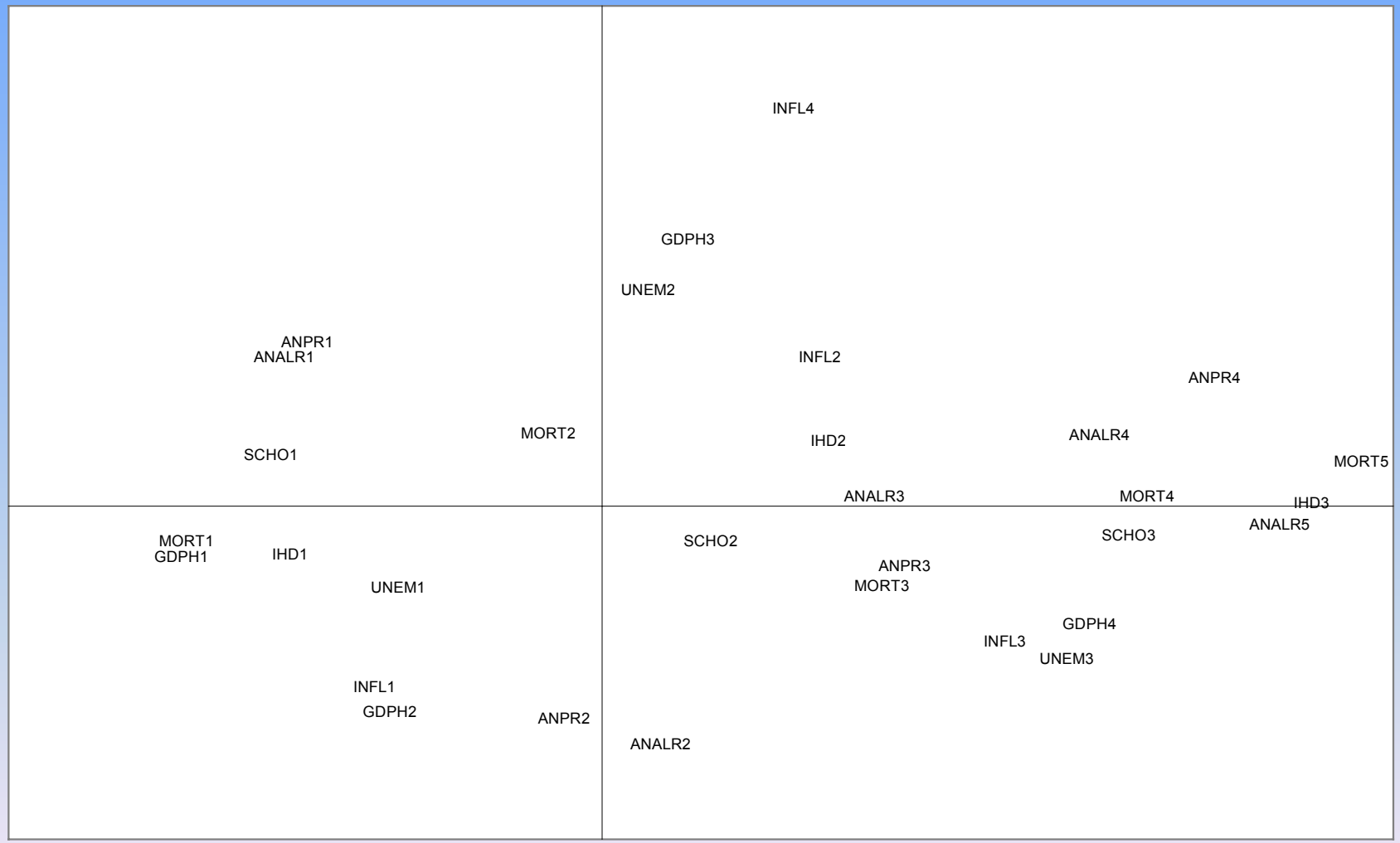
La répartition des 31 modalités sur la carte de Kohonen avec KACM, 500 itérations

ANPR2 UNEM2		MORT2 ANALR2	SCHO2		MORT3 ANALR3
	GDPH2		GDPH3	INFL2 IHD2	ANPR3
UNEM1		INFL3	INFL4		UNEM3
IHD1	INFL1		ANALR4	ANPR4	
ANPR1 ANALR1 SCHO1			MORT4	IHD3	MORT5 ANALR5
	MORT1 GDPH1		GDPH4	FSCHO3	

La représentation ACM, axes 1 (24%), et 2 (14%) 38% d'inertie expliquée



La représentation ACM, axes 1 (24%), et 5 (6%)
30% d'inertie expliquée



Macro-classes regroupant les modalités en 6 classes facilement interprétables

ANPR2 UNEM2		MORT2 ANALR2	SCHO2		MORT3 ANALR3
	GDPH2		GDPH3	INFL2 IHD2	ANPR3
UNEM1		INFL3	INFL4		UNEM3
IHD1	INFL1		ANALR4	ANPR4	
ANPR1 ANALR1 SCHO1			MORT4	IHD3	MORT5 ANALR5
	MORT1 GDPH1		GDPH4	FSCHO3	

Analyse des variables qualitatives en gardant les individus : KACM_ind

Il peut être intéressant de regrouper ensemble les individus et les modalités qui les décrivent. Dans ce cas, il faut utiliser le Tableau Disjonctif Complet afin de connaître les réponses individuellement.

Pour représenter simultanément les individus et les modalités, on construit une carte de Kohonen avec les individus en utilisant l'algorithme SOM appliqué au Tableau Disjonctif Complet Corrigé et on projette les modalités (normalisées) comme des données supplémentaires.

Chaque modalité *j* est représentée par un *M*-vecteur, qui est le *vecteur moyen* de tous les individus partageant cette modalité.

Ses coordonnées sont:

$$\frac{b_{jl}}{b_j\sqrt{b_l}\sqrt{K}}, \quad \text{for} \qquad l = 1, \cdots, M,$$

Chaque vecteur moyen est affecté à la classe de Kohonen de son plus proche code vecteur.

Cette méthode est nommée KACM_ind et fournit une représentation simultanée des individus et des modalités.

Représentation des 96 pays et des 31 modalités, 20 000 itérations, 7 macro-classes

INFL4 Moldavia Romania Russia Ukraine	Bulgaria Poland	GDPH3 Costa Rica Ecuador Jamaica Lebanon	Colombia Fiji Panama Peru Thailand		MORT5 Afghanistan Angola Haiti Mozambique Pakistan Yemen
Brazil	ANPR2 MORT2	Croatia Venezuela	ANALR2 UNEM2	Ghana Mauritania Sudan	ANPR4 ANALR5 IHD3
GPDH2 Chili Cyprus S. Korea	Argentina Bahrain Malaysia Malta Mexico		INFL3 Macedonia Mongolia	SCHO3 GPDH4 UNEM3 Laos	MORT4 ANALR4 Cameroon Comoros Ivory Coast Nigeria
Greece Hungary Slovenia Uruguay	UNEM1 IHD1 Portugal	China Philippines Yugoslavia	Albania Indonesia Sri Lanka	INFL2 Guyana Vietnam	Bolivia Kenya Nicaragua
ANPR1 ANALR1 SCHO1 R. Czech	Germany, Sweden Australia, Israel USA, Iceland Japan, Norway New-Zealand Netherlands United Kingdom Switzerland	INFL1 U Arab Emirates	SCHO2 Egypt	ANPR3 IHD2 Morocco Paraguay	ANALR3 El Salvador Swaziland
Belgium Canada Denmark Spain, Italy Finland France Ireland	MORT1 GDPH1 Luxemburg	Singapore	Saudi Arabia Syria	MORT3 Algeria	South Africa Iran Namibia Tunisia Turkey Zimbabwe

Un nouvel algorithme pour l'analyse simultanée des individus et des modalités : KDISJ

On utilise Le Tableau Disjonctif Complet corrigé D^c.
On choisit un réseau de Kohonen.
On associe à chaque unité u du réseau un vecteur code C_u formé de $(M + N)$ composantes, les M premières évoluent dans l'espace des individus (représentés par les lignes de D^c), les N dernières dans l'espace des modalités (représentées par les colonnes de D^c).

Pour mettre en évidence la structure du vecteur code C_u, on note

$$C_u = (C_M, C_N)_u = (C_{M,u}, C_{N,u})$$

Les étapes de l'apprentissage du réseau de Kohonen sont doubles.

On tire alternativement une ligne de D^c (c'est-à-dire un individu *i)*, puis une colonne (c'est-à-dire une modalité *j)*.

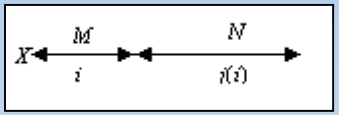
KDISJ (suite)

Quand on tire un individu i, on lui associe la modalité j (i) définie par

$$j(i) = Arg\max_j d_{ij}^c = Arg\max_j \frac{d_{ij}}{\sqrt{Kd_{.j}}}$$

C'est sa modalité la plus rare.

Ensuite, on crée un vecteur individu étendu $X = (i, j(i)) = (X_M, X_N)$, de dimension $(M + N)$.



On cherche alors parmi les vecteurs codes celui qui est le plus proche, au sens de la distance euclidienne restreinte aux M premières composantes. Notons u_0 l'unité gagnante. On rapproche alors les vecteurs codes de l'unité u_0 et de ses voisines du vecteur complété $(i, j(i))$, selon la loi usuelle de Kohonen.

KDISJ (suite)

On peut écrire formellement cette étape ainsi :

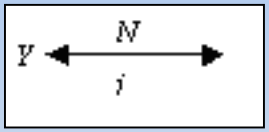
$$\begin{cases} u_0 = Arg\ \min_u \left\| X_M - C_{M,u} \right\| \\ C_u^{new} = C_u^{old} + \varepsilon\, \sigma(u, u_0)(X - C_u^{old}) \end{cases}$$

où ε est le paramètre d'adaptation (positif, décroissant avec le temps), et σ est la fonction de voisinage, avec $\sigma(u, u_0)$= 1 si u et u_0 sont voisins dans le réseau de Kohonen, et = 0 sinon. Le rayon de voisinage est aussi une fonction décroissante du temps.

KDISJ (fin)

Quand on tire une modalité j, de dimension N (une colonne de D^c), on ne lui associe pas d'individu. En effet, par construction, il y a beaucoup d'individus ex-æquo et le choix serait arbitraire.
On cherche parmi les vecteurs codes celui qui est le plus proche, au sens de la distance euclidienne restreinte aux N dernières composantes. Soit v_0 l'unité gagnante. On rapproche alors les N dernières composantes du vecteur code gagnant associé à v_0 et de ses voisins de celles du vecteur modalité j, sans modifier les M premières composantes.

Pour simplifier, notons Y le vecteur colonne de dimension N correspondant à la modalité j.



Cette étape peut s'écrire :

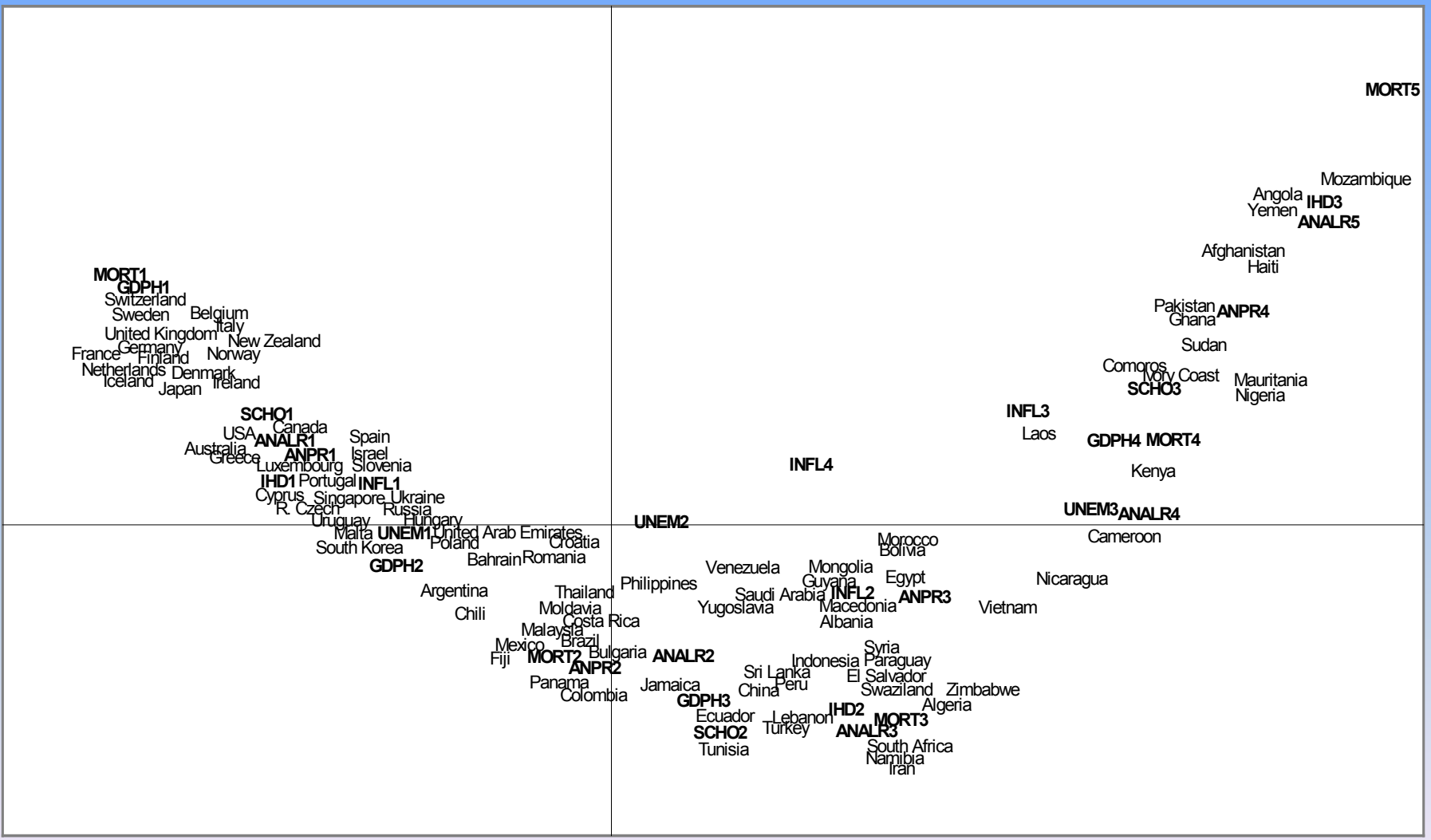
$$\begin{cases} v_0 = Arg\ \min_u \left\| Y - C_{N,u} \right\| \\ C_{N,u}^{new} = C_{N,u}^{old} + \varepsilon\, \sigma(u, v_0)(Y - C_{N,u}^{old}) \end{cases}$$

KDISJ pour les 96 pays et les 31 modalités, 2400 itérations et 7 macro-classes

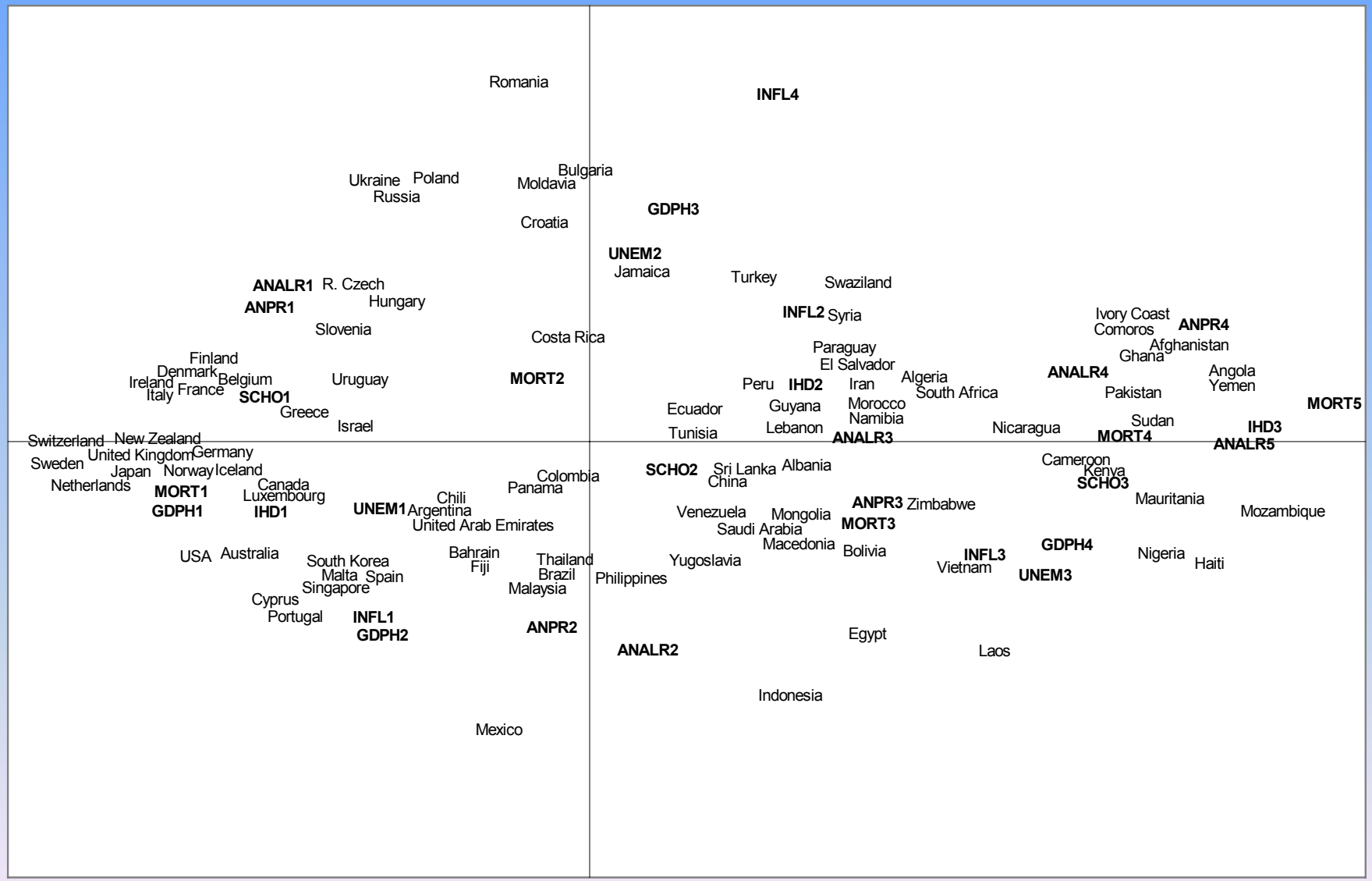
SCHO2 IHD2 Algeria Syria	Saudi Arabia Egypt Indonesia	Brazil Mexico	Argentina Chili Cyprus S. Korea	INFL1 Australia Canada USA	GDPH1 Belgium Denmark Finland France Ireland Italy
ANALR3 South Africa Iran Namibia El Salvador Zimbabwe	MORT3 Guyana Morocco Paraguay Tunisia Turkey		GDPH2	IHD1 Israel	MORT1 Germany, U Kingdom Iceland, Japan Lux, Singapore Norway, Sweden New Zealand Netherlands, Spain Switzerland
Kenya Nicaragua	Swaziland	INFL2 U Arab Emirates Malaysia UNEM2	Malta Portugal	UNEM1 Greece Hungary Slovenia Uruguay	ANPR1 ANALR1 SCHO1 R. Czech
ANALR4 Comoros Ivory Coast	MORT4 Cameroon Nigeria	ANPR3 Bolivia	Bahrain Philippines	Poland	INFL4 Croatia Moldavia Romania Russia Ukraine
ANPR4 IHD3 Ghana	SCHO3 GDPH4 Laos Mauritania Sudan	UNEM3 Vietnam Yugoslavia		MORT2 ANALR2 Bulgaria Ecuador Jamaica Lebanon, Peru	GDPH3 Costa Rica
MORT5 ANALR5 Afghanistan Angola Pakistan Yemen	Haiti Mozambique	INFL3 Macedonia Mongolia	Albania China Sri Lanka	Colombia Panama	ANPR2 Fiji Thailand Venezuela

Représentation ACM (modalités et individus), axes 1 (24%) et 2 (14%)

38% d'inertie expliquée



Représentation ACM (modalités et individus), axes 1 (24%) et 5 (6%)
30% d'inertie expliquée



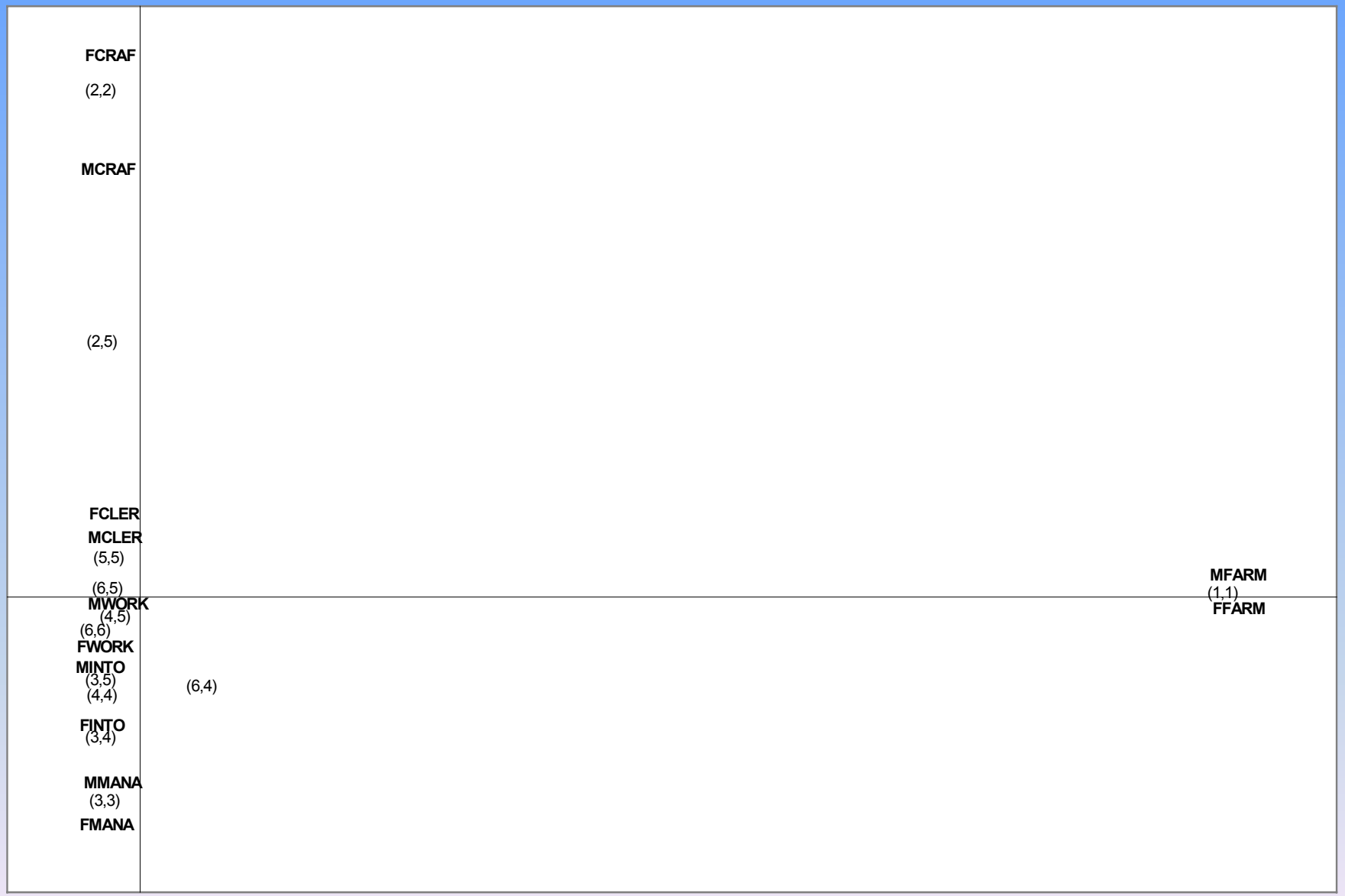
Un exemple jouet, les mariages

- On considère 270 couples, le mari et la femme étant classés dans 6 catégories professionnelles : agriculteur (farmer), artisan (craftsman), cadre (manager), profession intermédiaire (intermediate occupation), employé (clerk), ouvrier (worker) ; ces 6 catégories sont numérotées de 1 à 6.
- Ces données sont particulièrement simples puisque le tableau de contingence a une diagonale principale très dominante. On sait que la plupart des mariages se font à l'intérieur d'un même groupe professionnel. Le tableau disjonctif complet n'est pas montré mais il est très simple à calculer puisqu'il n'y a que deux variables (la catégorie professionnelle du mari et celle de la femme). Parmi les 36 combinaisons possibles de couples, seulement 12 sont présentes. Ainsi cet exemple, bien que construit à partir de données réelles, peut être considéré comme un exemple jouet.

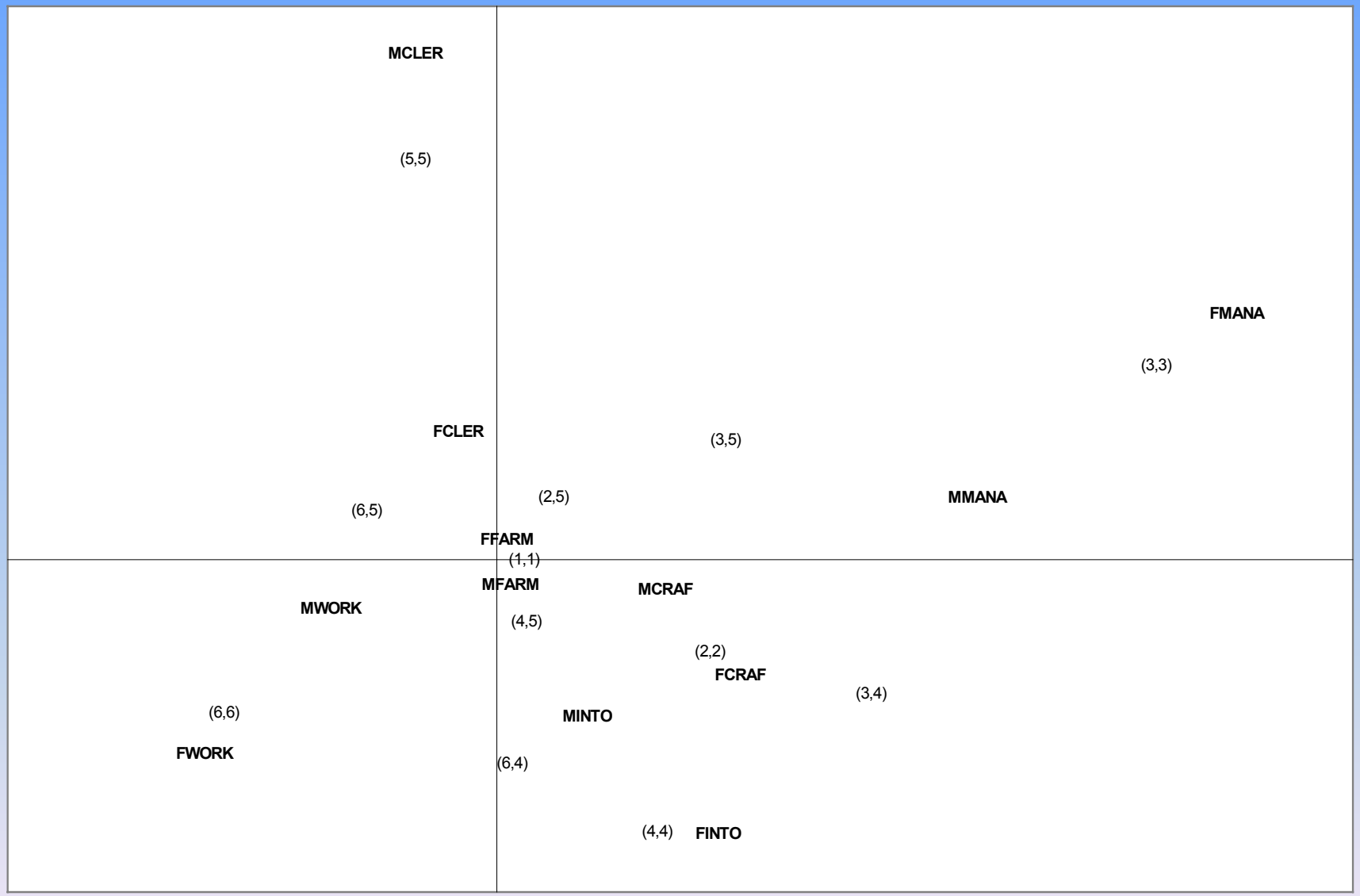
Tableau de contingence pour des couples mariés, (source INSE, 1990). Les lignes sont pour les hommes et les colonnes pour les femmes

	FFARM	FCRAF	FMANA	FINTO	FCLER	FWORK	Total
MFARM	16	0	0	0	0	0	16
MCRAF	0	15	0	0	12	0	37
MMANA	0	0	13	15	12	0	40
MINTO	0	0	0	25	35	0	60
MCLER	0	0	0	0	25	0	25
MWORK	0	0	0	10	60	32	102
Total	16	15	13	50	144	32	270

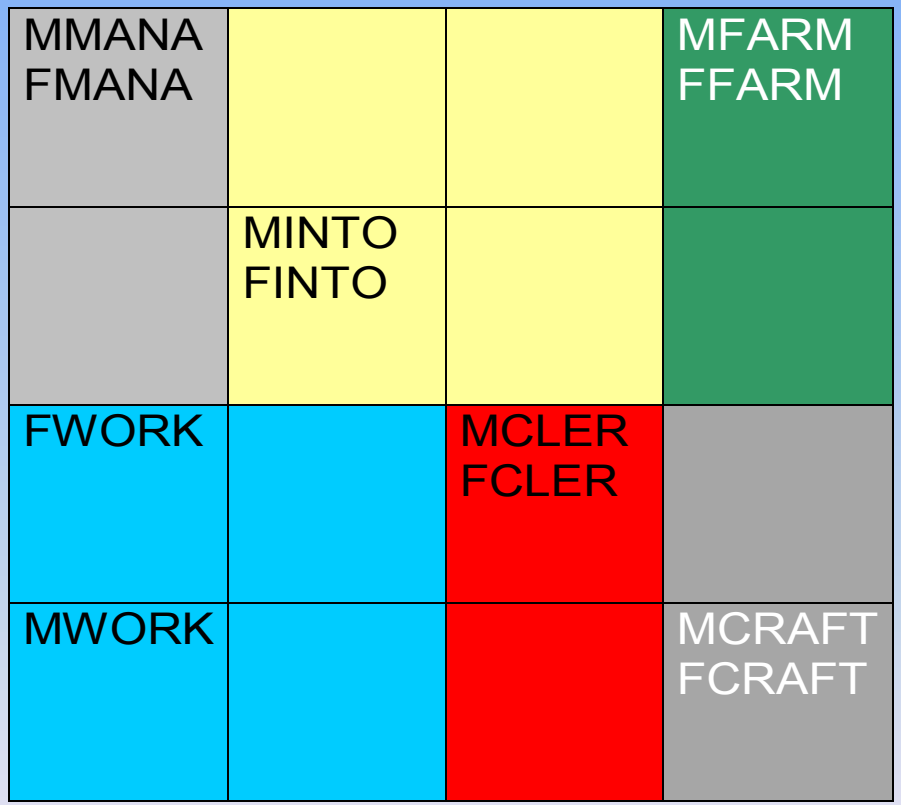
L’ACM (modalités et individus), axes 1 (20%) et 2 (17%), 37% d’inertie expliquée



L’ACM (modalités et individus), axes 3 (16%) et 5 (13%), 29% d’inertie expliquée



KACM sur les 12 modalités. Les 16 micro-classes sont regroupées en 6 macro-classes, 200 itérations



KACM_ind (modalités et individus). Le nombre de couples de chaque type est indiqué, 10000 itérations

MFARM FFARM 16 (1,1)		FMANA 13 (3,3)	MMANA
	MCLER 25 (5,5)		 15 (3,4) 12 (3,5)
MCRAFT FCRAFT 15 (2,2) 12 (2,5)	FCLER	 60 (6,5)	MWORK
	MINTO 25 (4,4) 35 (4,5)	FINTO 10 (6,4)	FWORK 32 (6,6)

KDISJ (modalités et individus). Le nombre de couples de chaque type est indiqué, 5000 itérations

MFARM FFARM 16 (1,1)		FINTO 25 (4,4) 10 (6,4)	MMANA 15 (3,4)
	MINTO 35 (4,5)		FMANA 13 (3,3) 12 (3,5)
MCRAFT FCRAFT 15 (2,2) 12 (2,5)	FCLER	MWORK 60 (6,5)	
	MCLER 25 (5,5)		FWORK 32 (6,6)

Les emplois intérimaires

- Le dernier exemple est tiré d'une vaste étude du temps de travail en 1998-1999 faite par l'INSEE à partir d'une enquête.
- On cherche à étudier les « formes particulières d'emplois» (FPE), à savoir les CDD, les emplois à temps partiel, les emplois intérimaires.
- L'étude analyse quelles contraintes spécifiques subissent les travailleurs concernés par les FPE.
- Dans l'enquête, les salariés devaient répondre à des questions concernant : leur durée du travail, le rythme de travail, la régularité, la flexibilité, la prévisibilité …
- Ici, nous nous intéressons à 115 travailleurs intérimaires.
- Nous présentons une application des algorithmes KACM et KDISJ utilisés aussi bien pour classer les modalités que les individus.

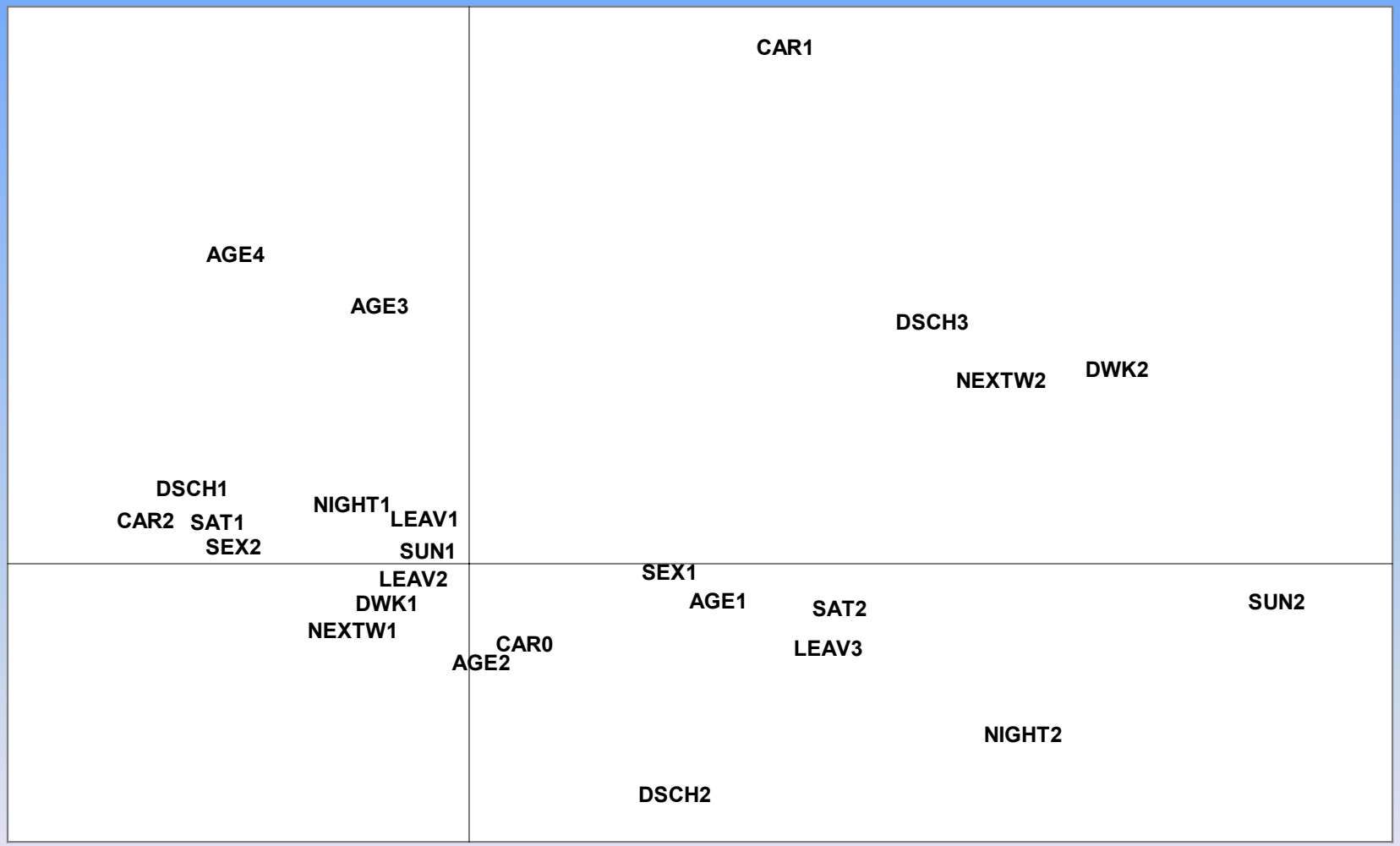
Ce tableau recense 10 variables (questions) et ses 25 modalités de réponse

Heading	Name	Response modalities
Sex	**Sex 1 2**	Man, Woman
Age	**Age 1, 2, 3, 4**	<25, [25, 40[, [40,50[, ≥50
Daily work schedules	**Dsch 1, 2, 3**	Identical, as-Posted, Variable
Number of days worked in a week	**Dwk 1, 2**	Identical, Variable
Night work	**Night 1, 2**	No, Yes
Saturday work	**Sat 1, 2**	No, Yes
Sunday work	**Sun 1, 2**	No, Yes
Ability to go on leave	**Leav 1, 2, 3**	Yes no problem, yes under conditions, no
Awareness of next week schedule	**Nextw 1, 2**	Yes, no
Possibility of carrying over credit hours	**Car 0, 1, 2**	No point, yes, no

KACM pour les 25 modalités regroupées en 6 clusters, 500 itérations

DWK2 SUN2		AGE4	CAR2	LEAV2
NEXWT2	DSCH3 CAR1		AGE3	SEX2
LEAV3		AGE2		DSCH1 SAT1
AGE1		SEX1	LEAV1	NIGHT1
DSCH2 NIGHT2 SAT2		CAR0		DWK1 SUN1 NEXTW1

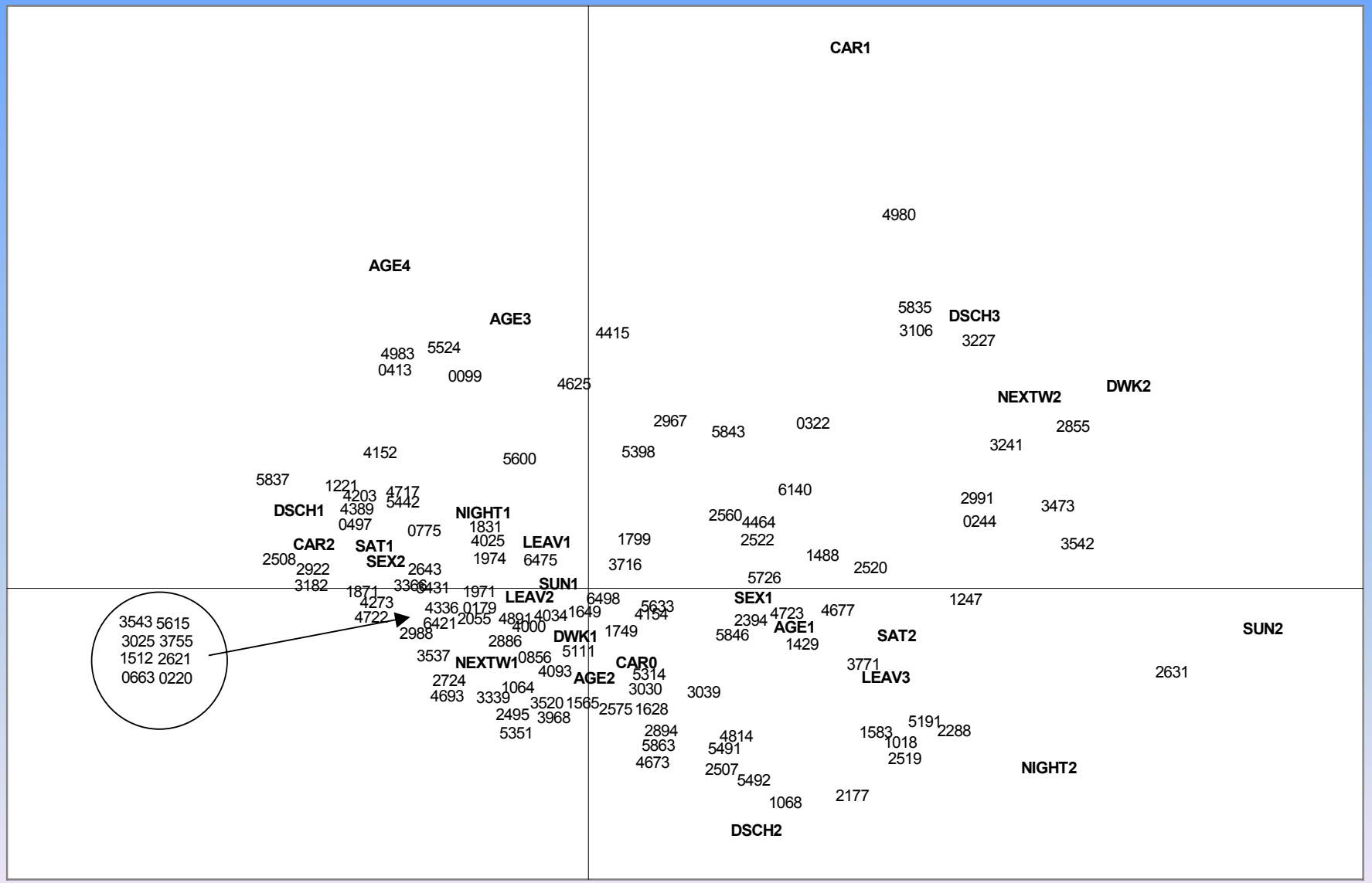
L’ACM (modalités), axes 1 (19%) et 2 (11%), 30% d’inertie expliquée



KDISJ (modalités et individus), seul le nombre d'individus est indiqué dans chaque classe, 3000 itérations

DWK2 6 ind	3 ind	DSCH3 7 ind	CAR1 1 ind	AGE3 9 ind
2 ind	NEXTW2 2 ind	LEAV3 3 ind	2 ind	4 ind
SUN2 6 ind	2 ind	4 ind	5 ind	LEAV2 6 ind
4 ind	SEX1 DSCH2 CAR0 4 ind	AGE2 DWK1 SUN1 NEXTW1 7 ind	SEX2 DSCH1 NIGHT1 SAT1 3 ind	CAR2 7 ind
NIGHT2 SAT2 6 ind	AGE1 9 ind	LEAV1 5 ind	4 ind	AGE4 4 ind

L’ACM (modalités & individus), axes 1 (19%) et 2 (11%), 30% d’inertie expliquée



Conclusion

Dans le tableau qui suit, nous avons résumé les relations entre les algorithmes factoriels classiques et les algorithmes fondés sur SOM

Algorithmes fondés sur SOM	**Méthodes Factorielles**
algorithme SOM sur les lignes de la matrice X	ACP, diagonalisation sur $X'X$.
KACM (classification des modalités): algorithme SOM sur les lignes de B^c	ACM, diagonalisation sur $B^{c\prime}B^c$.
KACM_ind (classification des individus): algorithme SOM sur les lignes de D^c et placement des modalités en supplémentaires	ACM avec individus, diagonalisation de $D^{c\prime}D^c$ et de $D^cD^{c\prime}$.
KDISJ apprentissage couplé des lignes (individus) et des colonnes (modalités): SOM sur les lignes et les colonnes de D^c	

Conclusion (fin)

- En fait, pour les applications, il est nécessaire de combiner différentes techniques.
- Par exemple, lorsque les variables sont quantitatives, il est souvent intéressant de réduir d'abord la dimension en utilisant une ACP et de ne conserver qu'un nombre réduit de coordonnées.
- S'il y a à la fois des variables qualitatives et quantitatives, on peut construire une classification des observations en ne retenant que les variables quantitatives, en utilisant une classification de KOHONEN suivie d'une CAH afin de définir une nouvelle variabl qualitative. Elle vient s'ajouter aux autres variables qualitatives et on peut appliquer une ACM ou un KACM à toutes les variables qualitatives (les variables de départ et la nouvelle variable). Cette technique permet une interprétation des classes et montre la proximité entre les modalités.
- Si l'on s'intéresse aux seuls individus, on peut transformer les variables qualitatives en variables quantitatives grâce à une ACM, auquel cas on garde tous les axes et on peut alors décrire chaque observation par ses coordonnées. La base de données devient alors quantitative et on peut l'analyser grâce à un algorithme classique de classification ou pa un algorithme de Kohonen.
- Il serait utile d'avoir en tête toutes ces techniques, ainsi que les techniques classiques, afin d'améliorer leurs performances, et de les considérer comme des outils utiles dans l data mining.