

L'Algorithme de Kohonen pour l'analyse et la visualisation des données

Patrick Letremy et Marie Cottrell

SAMOS-MATISSE

CNRS UMR 8595

Université Paris 1- Sorbonne

Analyse de données : introduction

Algorithme de Kohonen

Kohonen et classification : KACP

Accélération de la classification

Traitements des variables qualitatives

Conclusion

Analyse de données : introduction

Algorithme de Kohonen

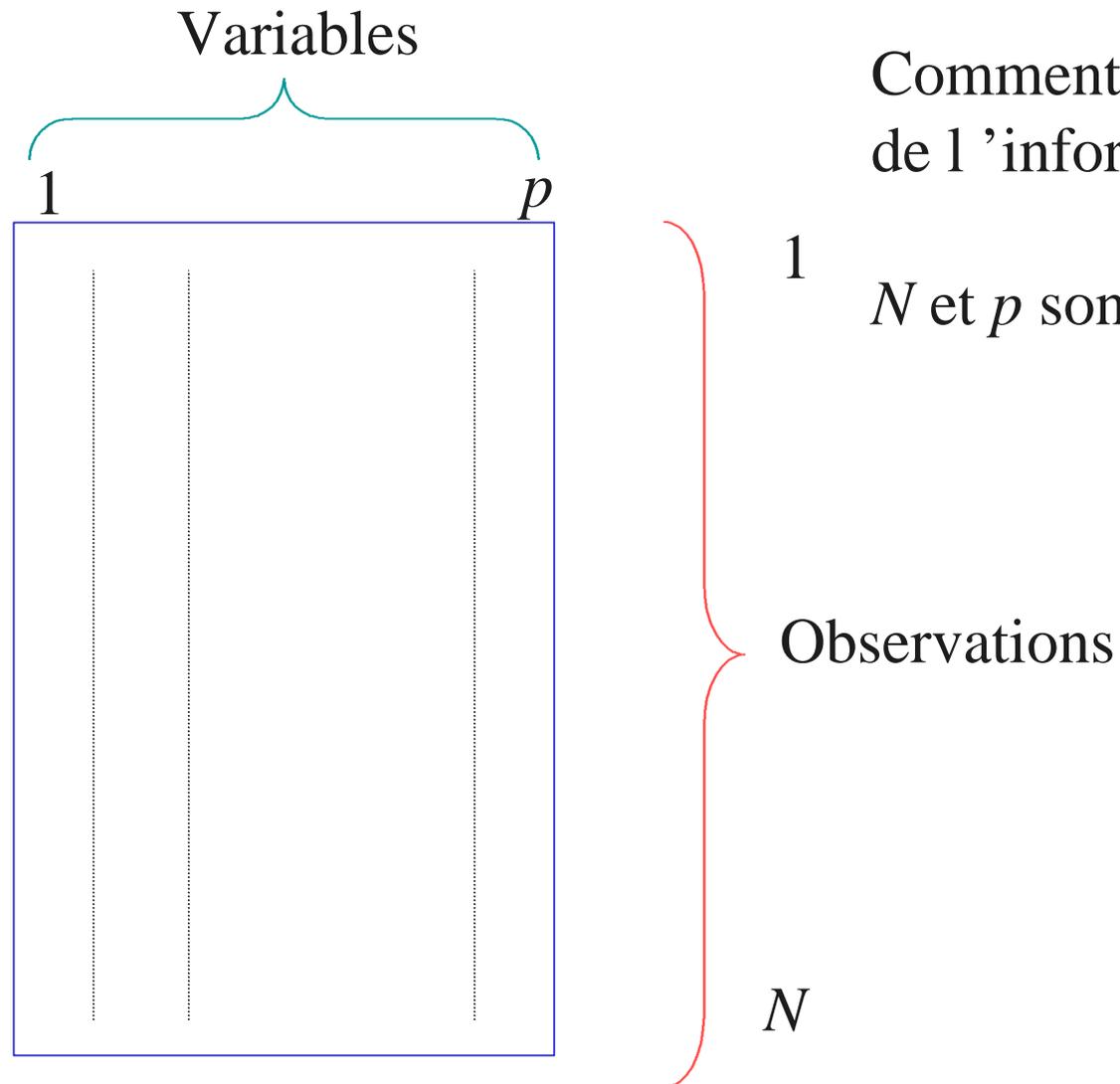
Kohonen et classification : KACP

Accélération de la classification

Traitements des variables qualitatives

Conclusion

Analyse de données



Comment extraire
de l'information ?

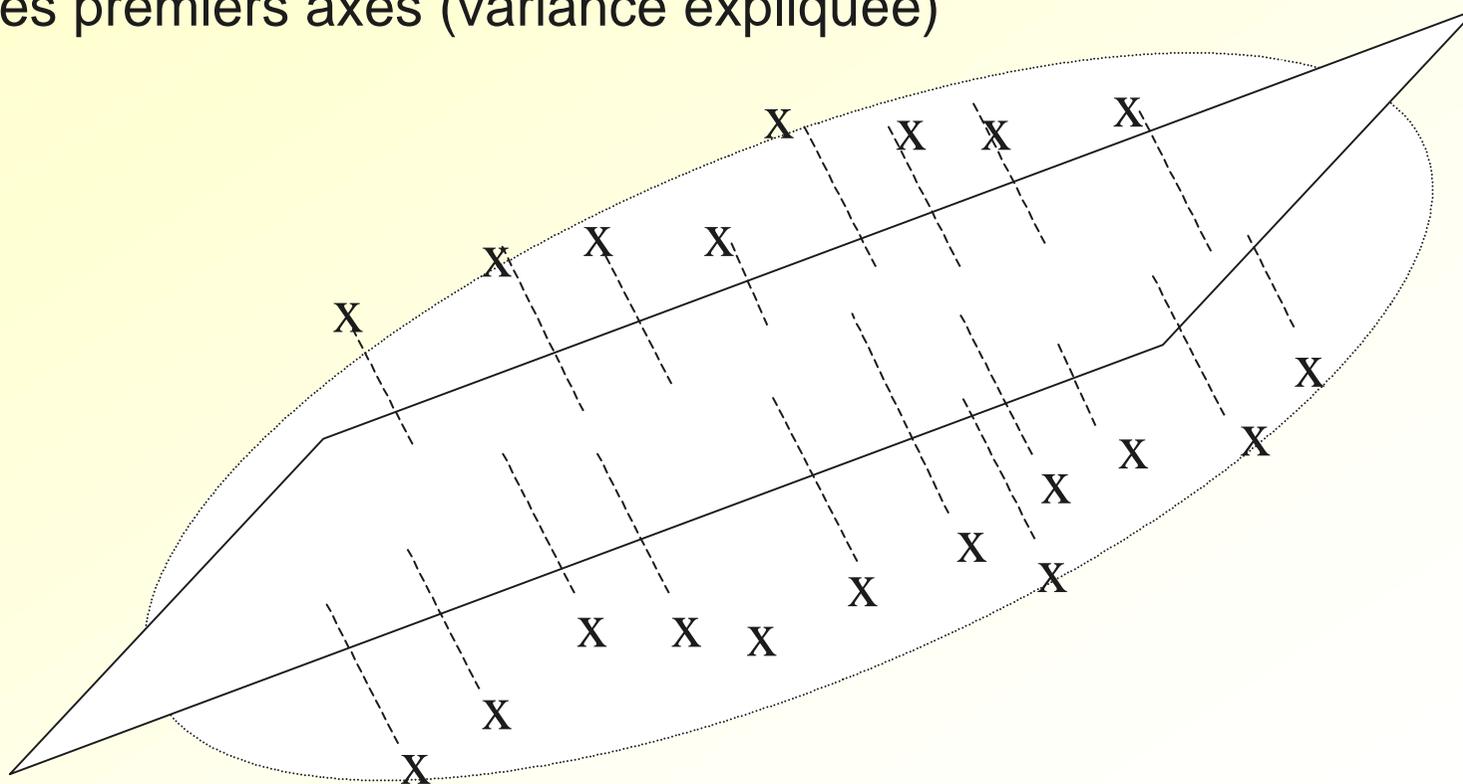
1
 N et p sont grands

Observations

N

Réduction du nombre de dimensions

- ✓ **L'analyse en composantes principales** consiste à changer d'axes dans l'espace des observations, de façon à maximiser la dispersion des projections sur les axes successifs.
- ✓ On peut donc se restreindre aux coordonnées des projections sur les premiers axes (variance expliquée)



Extraction d 'individus types :

Quantification Vectorielle

- ✓ K : espace des données, dimension p
- ✓ f : densité des données
- ✓ x_1, x_2, \dots, x_N : les données
- ✓ n : nombre de classes
- ✓ C_1, C_2, \dots, C_n : quantifieurs ou vecteurs codes
- ✓ G_1, G_2, \dots, G_n : classes

BUT : Minimiser la **distorsion quadratique** (l'erreur)
(= **Somme des carrés intra**)

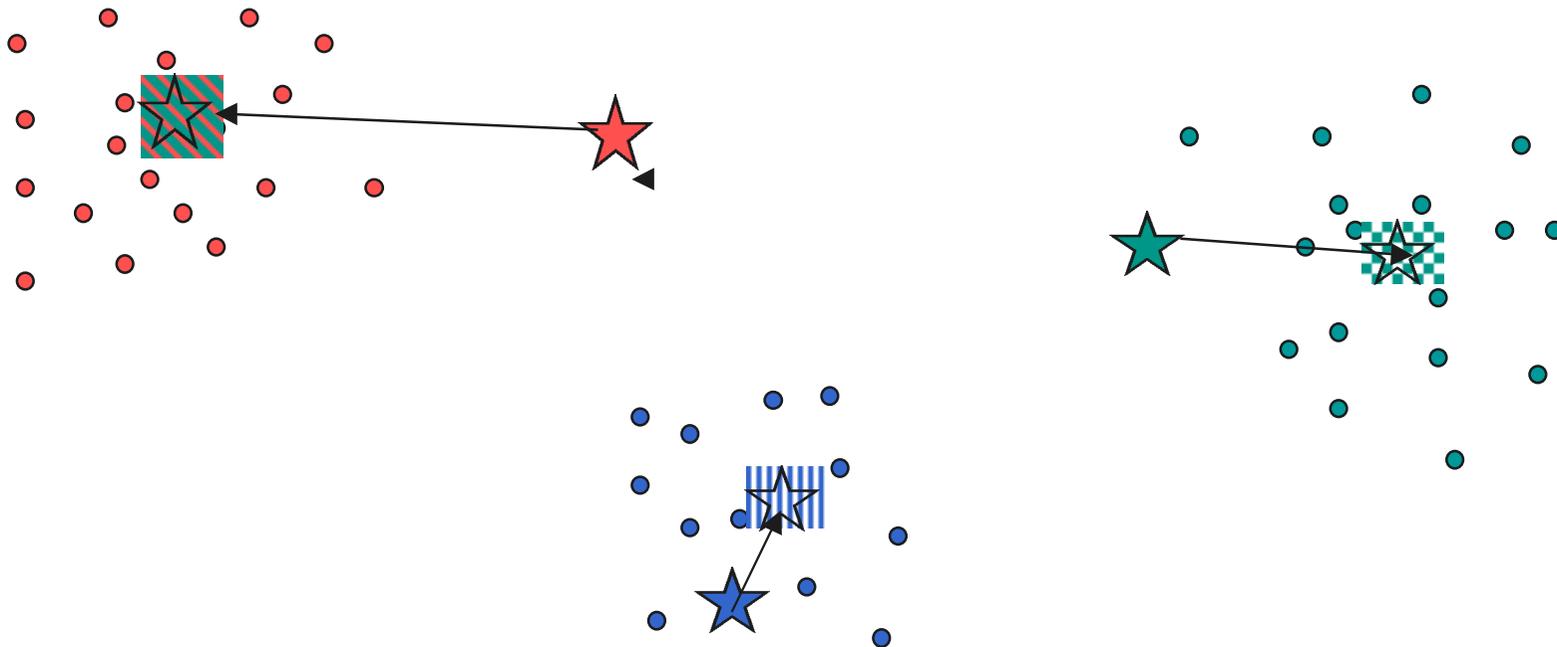
$$x_0(f, C_1, C_2, \dots, C_n) = \sum_{i=1}^n \int_{G_i} \|x - C_i\|^2 f(x) dx \quad (1)$$

Estimée par

$$\hat{x}_0(f, C_1, C_2, \dots, C_n) = \frac{1}{N} \sum_{i=1}^n \sum_{x_j \in G_i} \|x_j - C_i\|^2 \quad (2)$$

Algorithme Déterministe : Centres mobiles (FORGY, LLOYDS, LBG)

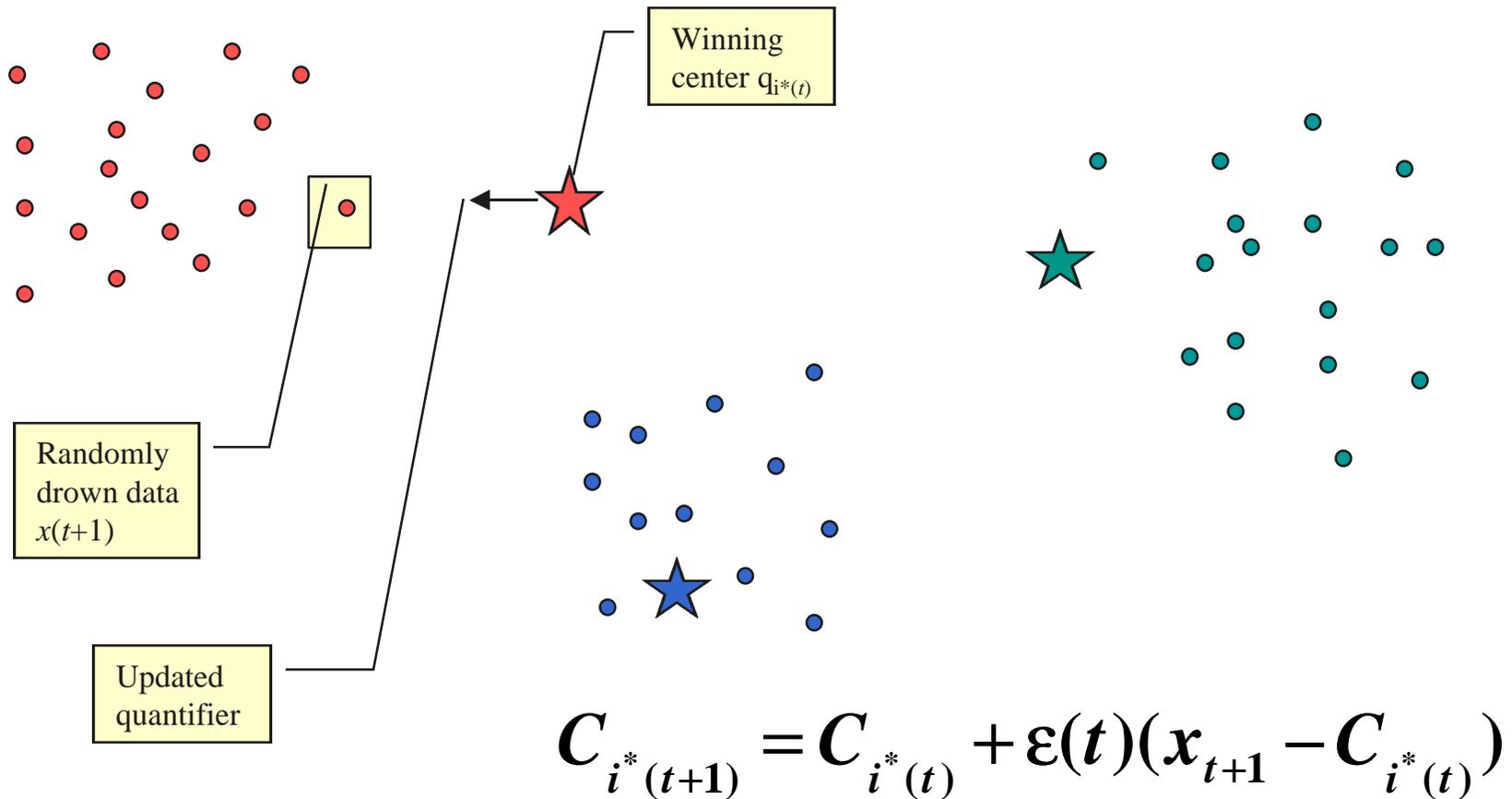
- ✓ A chaque étape, les classes sont définies (par les plus proches voisins), et les vecteurs codes sont re-calculés comme les centres de gravité des classes, etc.



- ✓ (On part de vecteurs codes aléatoires, on détermine les classes, puis les centres, puis les classes, etc.)

Algorithme Probabiliste associé (SCL)

- ✓ On déplace seulement le gagnant



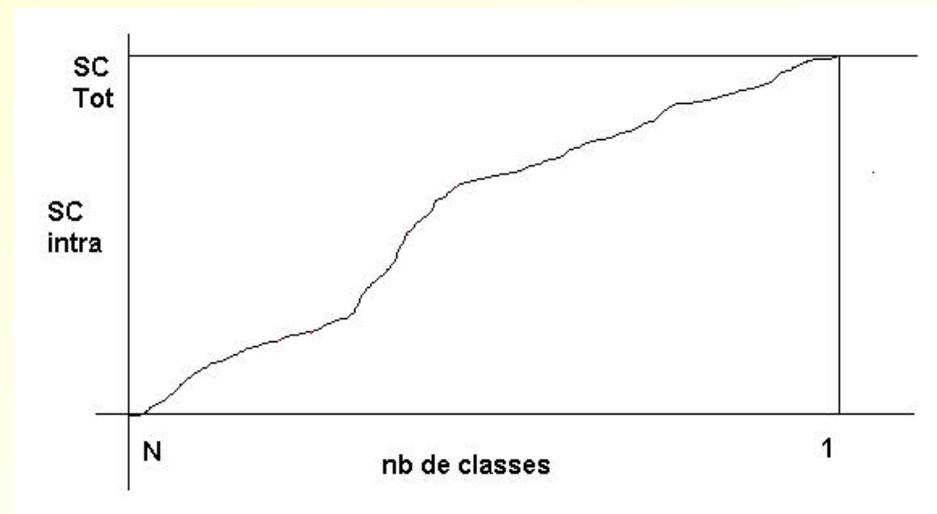
- ✓ *Avec l'algorithme de Kohonen, on déplace le vecteur code gagnant, mais aussi ses voisins.*

Algorithme SCL (0 voisin)

- ✓ L'algorithme SCL est la version stochastique de l'algorithme de Forgy
- ✓ L'algorithme de Forgy minimise la distorsion et converge vers un minimum local
- ✓ L'algorithme SCL converge ***en moyenne vers un minimum local***
- ✓ La solution dépend de l'initialisation

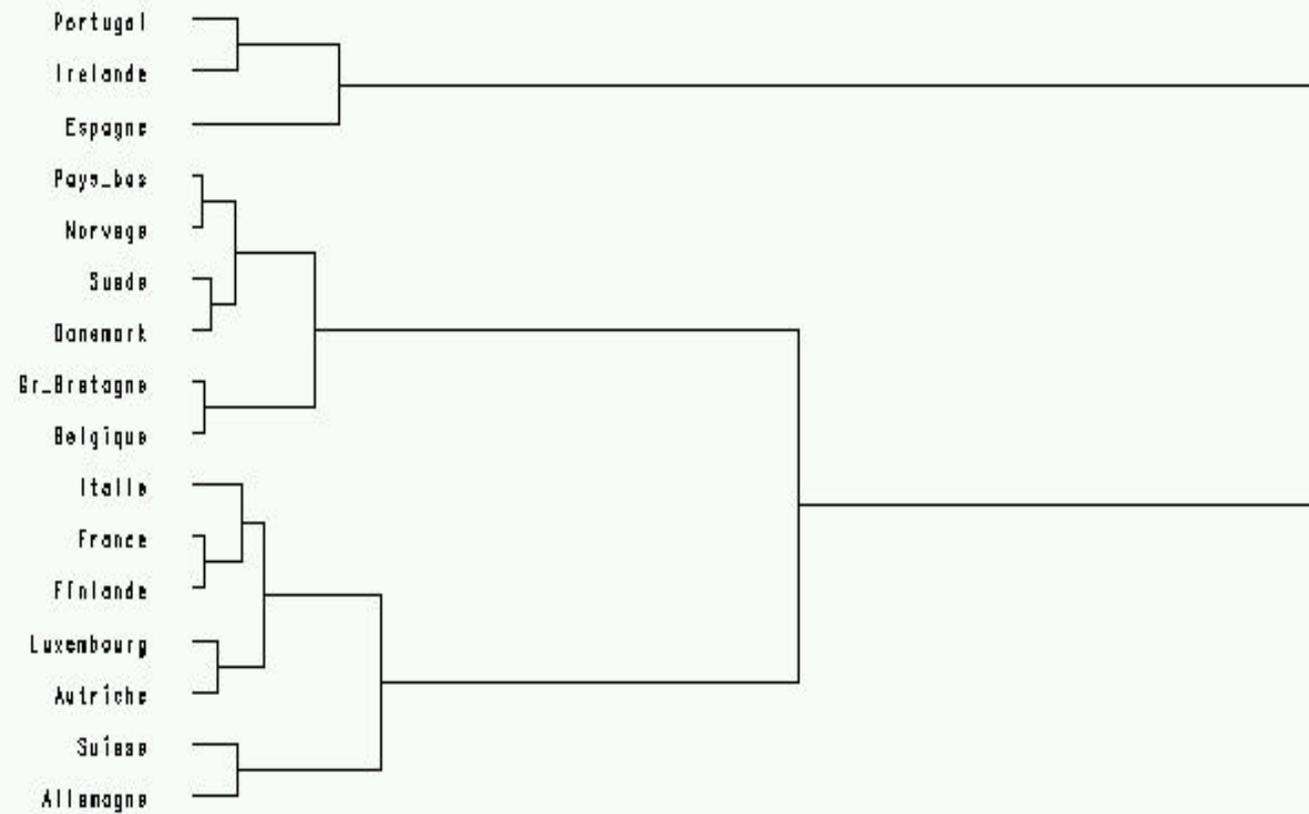
Classification Hiérarchique

- ✓ On construit une suite de classifications emboîtées en regroupant les individus les plus proches, puis les classes les plus proches, au sens d'une distance convenable
- ✓ Au fur et à mesure des regroupements, la somme des carrés intra classes augmente de 0 jusqu'à la somme des carrés totale
- ✓ On choisit en général la distance de Ward qui minimise à chaque étape les sauts de la somme des carrés intra



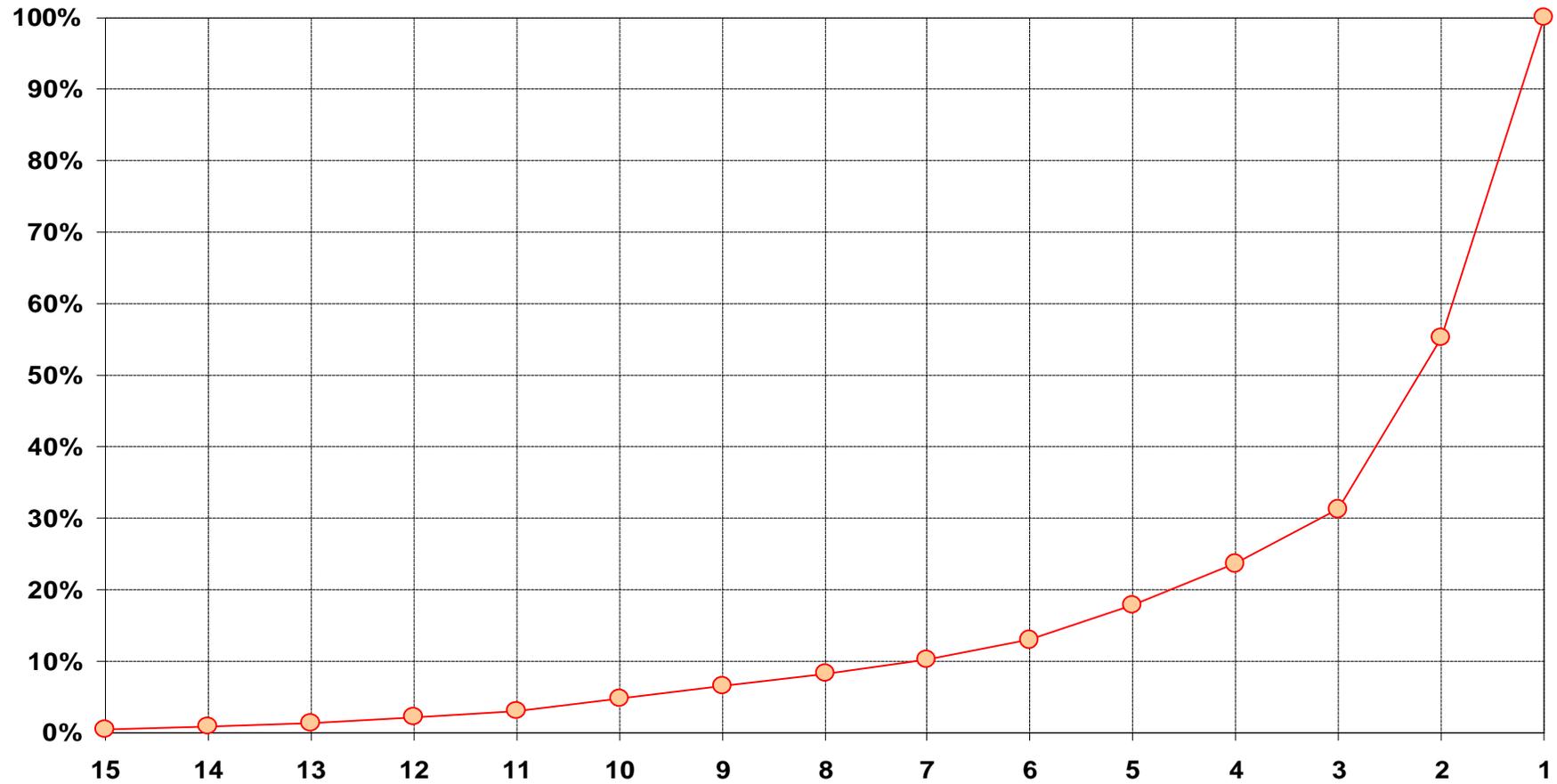
Arbre de classification

Exemple de Dendrogramme



Variation de la somme des carrés intra

INTRA/Totale



Nombre de classes décroissant de 15 à 1

Analyse de données : introduction

Algorithme de Kohonen

Kohonen et classification : KACP

Accélération de la classification

Traitements des variables qualitatives

Conclusion

Algorithme de Kohonen (SOM)

- ✓ Apprentissage non supervisé
- ✓ Les réponses associées à des entrées voisines sont voisines
- ✓ On parle d'auto-organisation, de respect de la topologie
- ✓ Les associations
 - rétine - cortex visuel
 - fréquences des sons - cortex auditif
 - peau - cortex sensoriel

respectent la notion de voisinage

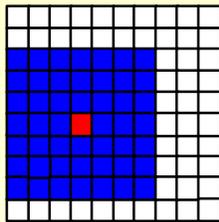
- ✓ Nombreuses applications en représentation de données de grande dimension sur des réseaux de dimension 1 ou 2, ou classification où la notion de classes voisines a un sens

L'algorithme

- ✓ Il s'agit d'un algorithme original de classification qui a été défini par Teuvo Kohonen, dans les années 80.
- ✓ L'algorithme regroupe les observations en classes, en respectant la topologie de l'espace des observations. Cela veut dire qu'on définit a priori une **notion de voisinage entre classes** et que des **observations voisines** dans l'espace des variables (de dimension p) appartiennent (après classement) à la **même classe ou à des classes voisines**.
- ✓ Les voisinages entre classes peuvent être choisis de manière variée, mais en général on suppose que les classes sont disposées sur une grille rectangulaire qui définit naturellement les voisins de chaque classe.
- ✓ Mais on peut choisir une autre topologie

Structure en grille ou en ficelle

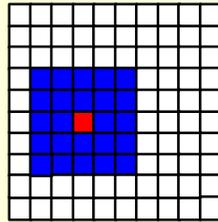
✓ Les grilles ne sont pas nécessairement carrées



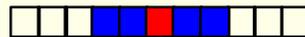
Voisinage de 49



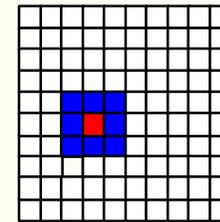
Voisinage de 7



Voisinage de 25



Voisinage de 5



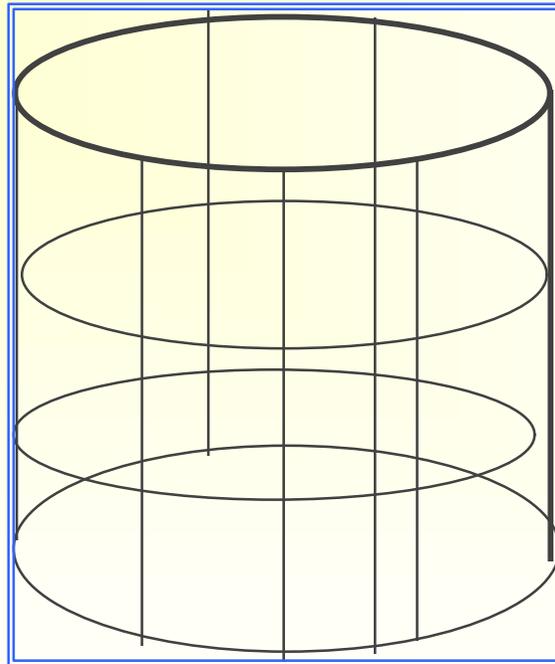
Voisinage de 9



Voisinage de 3

Structure en cylindre

CYLINDRE



L'algorithme

Principe de l'algorithme de Kohonen

- ✓ L'algorithme de **classement** est **itératif**. L'initialisation consiste à associer à chaque classe un vecteur code dans l'espace des observations choisi de manière aléatoire. Ensuite, à chaque étape, on choisit une observation au hasard, on la compare à tous les vecteurs codes, et on détermine la **classe gagnante**, c'est-à-dire celle dont le vecteur code est le plus proche au sens d'une distance donnée a priori. **On rapproche alors de l'observation les codes de la classe gagnante et des classes voisines.**
- ✓ Cet algorithme est analogue à **l'algorithme des centres mobiles**, mais dans ce dernier cas, il n'existe pas de notion de voisinage entre classes et on ne modifie à chaque étape que le code de la classe gagnante.
- ✓ C'est aussi un **algorithme compétitif**

Kohonen / SCL

- ✓ En fait l'algorithme de Kohonen est une extension de la version stochastique de l'algorithme des centres mobiles
- ✓ Issu du domaine de la quantification vectorielle, de la théorie du signal
- ✓ Applications où les données sont très nombreuses, disponibles on-line,
- ✓ Pas besoin de les stocker

Notations (Kohonen, ou SOM)

- ✓ Espace des entrées K dans R^p
- ✓ n unités, rangées en réseau de dimension 1 ou 2, pour lesquelles est défini un système de voisinage
- ✓ A chaque unité i ($i=1, \dots, n$), est associé un **vecteur code** C_i de p composantes

- ✓ La réponse d'une unité i à l'entrée x est mesurée par la proximité de x avec le vecteur poids C_i

- ✓ Initialisation aléatoire des poids
- ✓ A l'étape t ,
 - on présente une entrée x
 - on cherche l'unité gagnante $i_0(x)$
 - on rapproche C_{i_0} et les C_i voisins de l'entrée x

Définition de l'algorithme

✓ Cette modification s'écrit :

$$C_i(t+1) = C_i(t) + \varepsilon(x(t+1) - C_i(t))$$

où

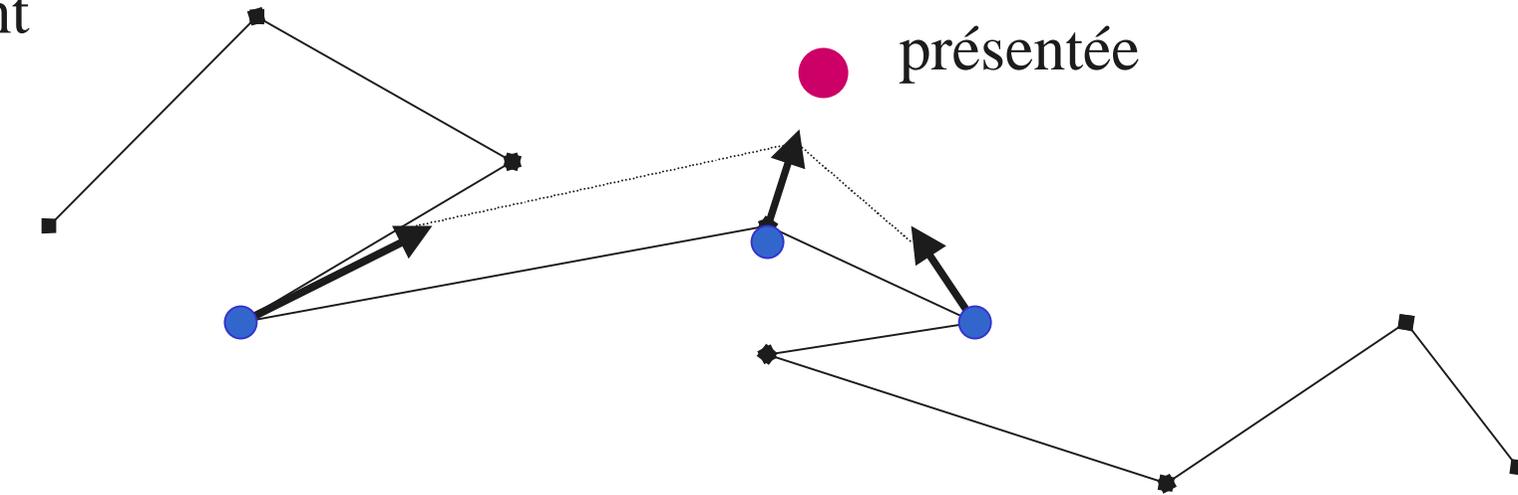
- ✓ $x(t+1)$ est l'observation présentée à l'étape $(t+1)$,
- ✓ C_i le vecteur code associé à la classe i ,
- ✓ ε un paramètre d'adaptation positif et inférieur à 1.

- ✓ Cette modification n'intervient que pour la classe gagnante $C_{i_0(x)}$ et ses voisines
- ✓ Deux paramètres : **le gain ε et le nombre de voisins**
ils décroissent lentement au cours du temps
- ✓ Les entrées sont présentées aléatoirement et indépendamment

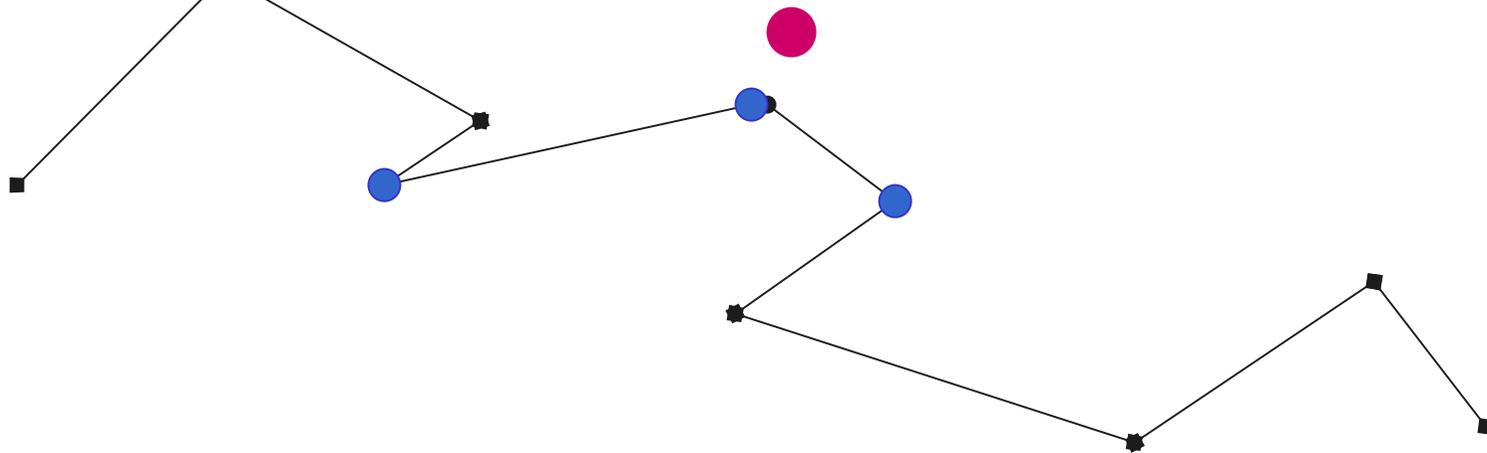
Exemple : une étape

Avant

Observation
présentée

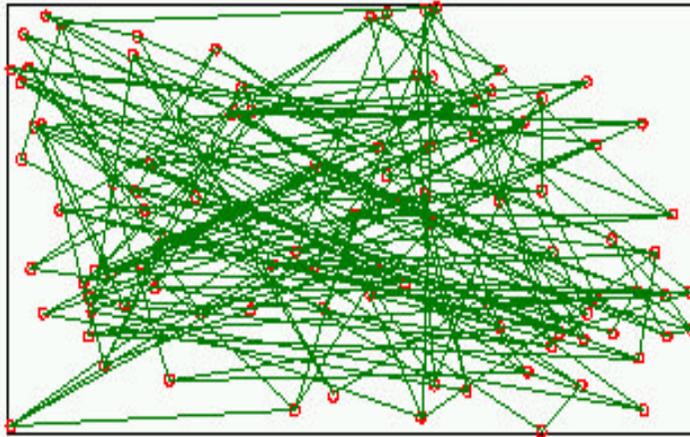


Après

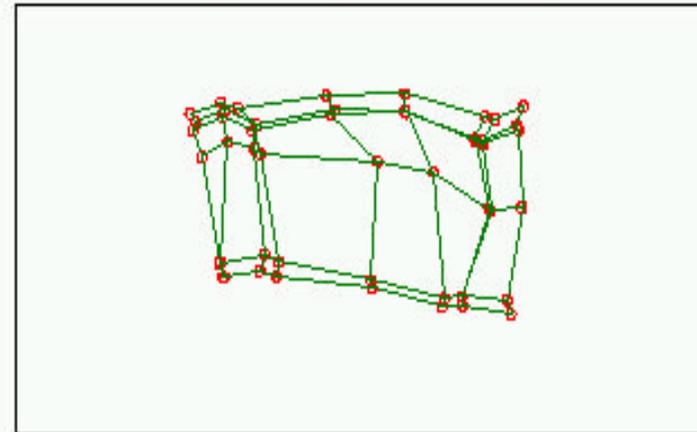


Démo en dimension 2

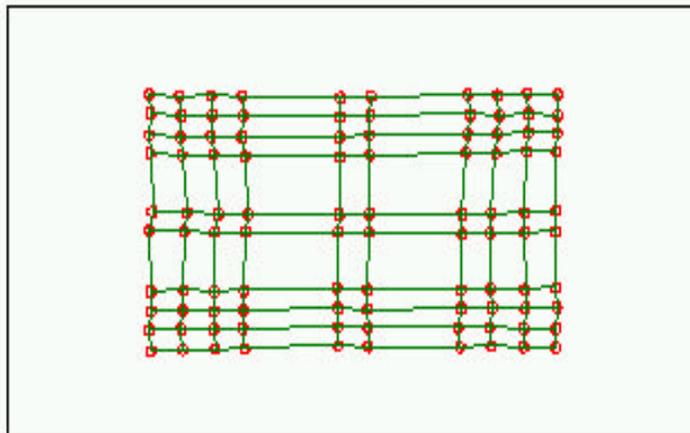
Grille:10x10 Etape= 0/1000000 Rayon= 0



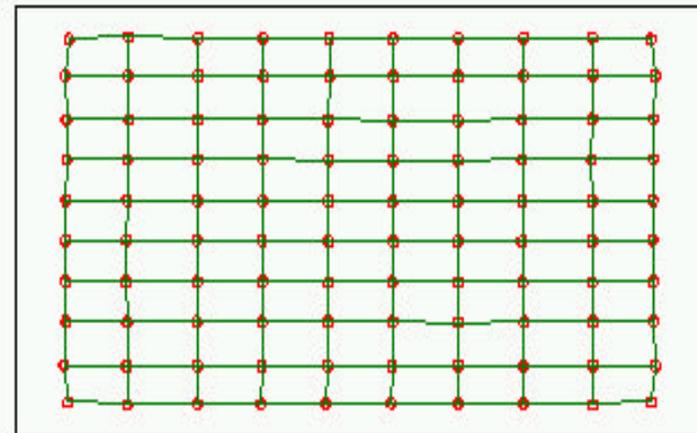
Grille:10x10 Etape= 1000/1000000 Rayon= 5



Grille:10x10 Etape= 100000/1000000 Rayon= 3



Grille:10x10 Etape= 1000000/1000000 Rayon= 0



Étude Théorique

- ✓ On peut écrire

$$C(t+1) = C(t) + e H(x(t+1), C(t))$$

- ✓ mais en général (si la distribution des entrées est continue), **H ne dérive pas d'un potentiel** (Erwinn)
- ✓ **On se restreint ici au cas où les entrées sont listées en nombre fini**
- ✓ Dans ce cas, l'algorithme minimise la somme des carrés des écarts de chaque observation non seulement à son vecteur code, mais aussi aux vecteurs codes voisins (dans la structure fixée)
- ✓ C'est une **extension de la notion de somme des carrés intra-classes, qui est étendue aux classes voisines**

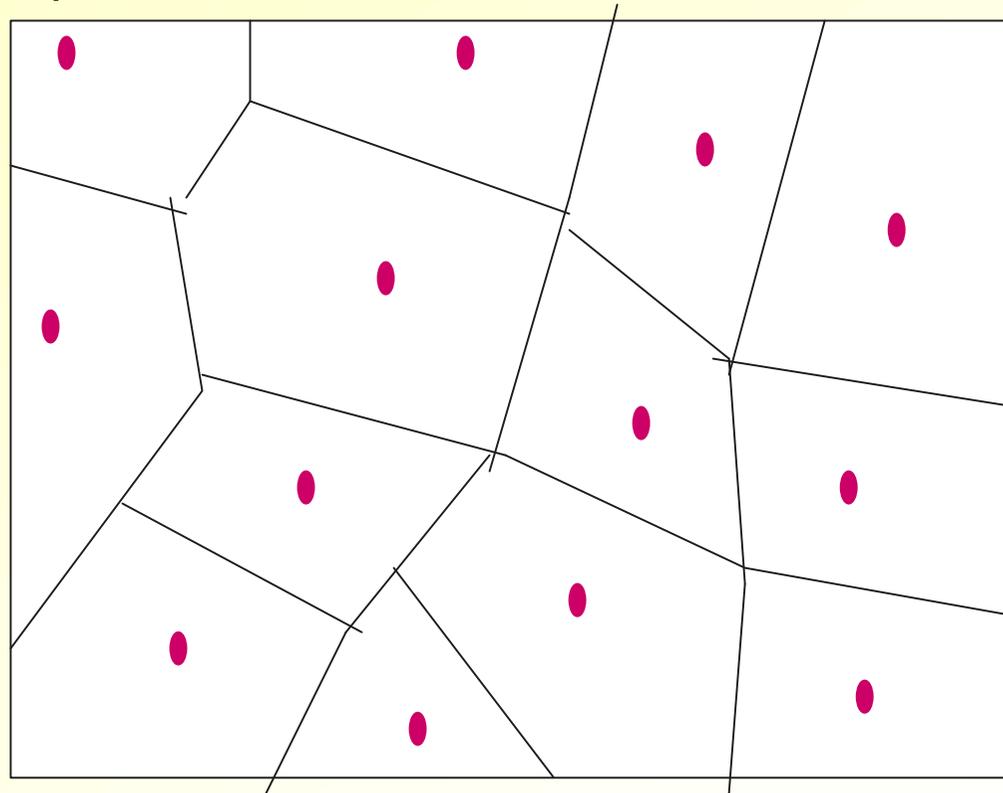
Somme des carrés intra-classes étendue aux classes voisines

$$V = \sum_{i=1}^n \sum_{x / i=i_0(x) \text{ ou } i \text{ voisin de } i_0(x)} \|x - C_i\|^2$$

- ✓ En fait cette fonction a de nombreux minima locaux
- ✓ L'algorithme converge, moyennant les hypothèses classiques (Robbins-Monro) sur les ε , qui doivent décroître ni trop, ni trop peu
- ✓ ***La démonstration mathématique complète n'est faite que pour des données de dimension 1 et pour une structure de voisinage en ficelle***
- ✓ Pour accélérer la convergence, on prend au début une taille de voisinage assez grande et on la fait décroître

Mosaïque de Voronoï

Dans l'espace des entrées, les classes forment une partition
Par exemple en dimension 2



X appartient à $G_i \Leftrightarrow$ l'unité i gagne quand on présente i

Analyse de données : introduction

Algorithme de Kohonen

Kohonen et classification : KACP

Accélération de la classification

Traitements des variables qualitatives

Conclusion

Cartes de Kohonen : Classification

- ✓ Pour représenter des données au moyen de l'algorithme de Kohonen, on prend comme entrées les lignes de la matrice des données
- ✓ Après apprentissage, chaque individu (ligne) correspond à une unité du réseau (celle qui gagne quand on présente cet individu)
- ✓ ***On classe une observation dans la classe G_i définie par l'unité gagnante qui lui correspond ($i=i_0(x)$)***
- ✓ On obtient donc une classification des individus, avec respect des voisinages
- ✓ La carte ainsi obtenue fournit une représentation plane
- ✓ Ici l'existence de proximités entre classes qui se ressemblent est essentielle

Représentation (KACP)

- ✓ Dans chaque classe on peut représenter le vecteur code
 - en donnant ses P composantes
 - en dessinant une courbe à P points
- ✓ Dans chaque classe, on peut
 - faire la liste des observations de cette classe
 - représenter en superposition les observations de la classe
- ✓ Ceci fournit une **représentation plane**, analogue à l'analyse en composantes principales (mais une seule carte et pas de projection orthogonale)

Nombreuses applications

- ✓ Représentation des pays, (Blayo et Letremy)
- ✓ Communes d'Ile-de France, (Ibbou, Tutin)
- ✓ Courbes de consommation
classification et prévision, (Rousset)
- ✓ Consommation au Canada, (Gaubert, Gardes, Rousset)
- ✓ Segmentation du marché du travail (Gaubert)
- ✓ Démographie et composition sociale dans la vallée du Rhône, (Letremy, P.A.R.I.S)
- ✓ Etude sur le rôle du leasing en Belgique, (de Bodt, Ibbou)
- ✓ Classification des profils de chocs de taux d'intérêts, (de Bodt), etc...

Un exemple : 96 pays en 1993

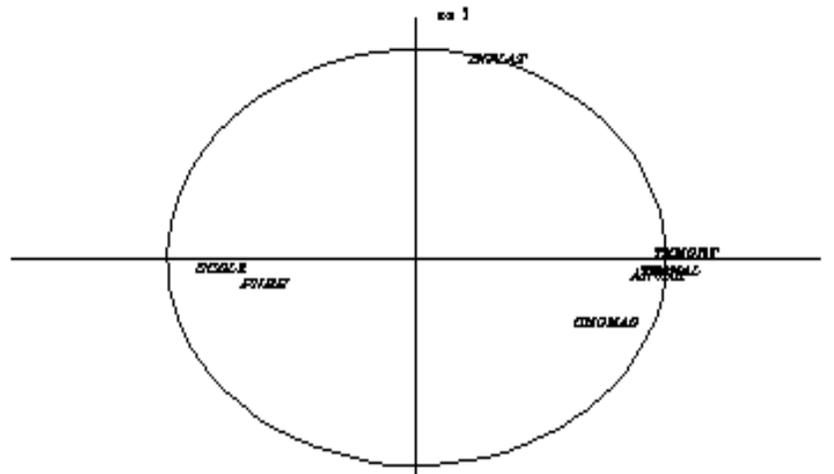
ANCRX	Croissance annuelle de la population en %
TXMORT	Taux de mortalité infantile (en pour mille)
TXANAL	Taux d'illettrisme en %
SCOL2	Indice de fréquentation scolaire au second degré
PNBH	PNB par habitant exprimé en dollars
CHOMAG	Taux de chômage en %
INFLAT	Taux d'inflation en %
NIVIDH	Niveau de l' Indice de Développement Humain (6 niveaux codés par libellés) (faible1, faible2, moyen1, moyen2, fort1, fort2)
CATIDH	Niveau d' Indice de Développement Humain (6 niveaux codés de 1 à 6)

Les données

PAYS	ANCRX	TXMORT	TXANAL	SCOL2	PNBH	CHOMAG	INFLAT	NIVIDH	CATIDH
Afghanistan	6	159	70,9	15	276	19	17	faible1	1
Afrique du sud	2,6	46,9	23,5	77	2873	33	10	moyen2	4
Albanie	1,1	33,1	8	29,2	828	17,9	16,1	moyen2	4
Algerie	2,2	42,1	42	61	1570	27	31	moyen2	4
Allemagne	0,2	5,8	1	101,2	24993	9,4	3,1	fort2	6
Angola	3,6	126,8	58	14	575	25	951	faible1	1
Arabie Saoudite	3	68,8	39,5	49	7081	6,6	0,7	moyen2	4
Argentine	1,1	33,8	4,4	72,3	7827	11,3	4	fort1	5
Australie	1,3	5,9	0,1	84	17688	9,7	2,5	fort2	6
Bahrein	2,5	24,2	17	99	7500	15	2	fort1	5
Belgique	0,1	7,8	0,3	103,2	22225	12,6	2,6	fort2	6
Bolivie	2,2	74,9	20	37	733	6,2	8,5	moyen1	3
Bresil	1,6	59,8	18	43	3073	5,5	1094	fort1	5
Bulgarie	-0,2	15,3	2,1	68,2	1058	17	33	moyen2	4
Cameroun	2,9	85,8	36,5	32	733	25,1	12,8	moyen1	3
Canada	1	6,7	3,1	104,2	18286	10,4	0,3	fort2	6
Chili	1,4	14,4	5,7	67	3643	6,1	11,2	fort1	5
Chine	1	25,8	22,4	55	418	2,5	22	moyen1	3
Chypre	1	9,9	4,5	95	9459	2	4,8	fort2	6
Colombie	1,7	36,8	8,5	62	1379	8	22,9	fort1	5
Comores	3,5	81,7	42,5	19	317	16	24,8	faible2	2
Coree du Sud	1	14,9	3,7	96	7572	2,3	6	fort1	5
Costa Rica	2,2	13,5	5,2	47	1896	5	15	fort1	5
Cote d' Ivoire	3,3	90,9	46,8	25	587	17	25,6	faible1	1
Croatie	0,1	11,5	3,2	83,2	2755	13,1	97,6	moyen2	4
Danemark	0,2	5,6	1	114,2	28346	12,1	2,1	fort2	6
Egypte	1,9	58,8	50,5	76	632	20,2	8,3	moyen1	3
Emirats arabes uni	2,2	23,2	20,9	89	23809	0,2	5	fort1	5
Equateur	2,1	36,8	12,8	55	1205	7,2	26	moyen2	4
Espagne	0,2	7,3	7,1	110,2	12283	24,4	4,8	fort2	6
Etats Unis	1	8,2	3	97,2	25219	5,6	2,8	fort2	6
Fidji	1,6	26,9	9	64	2077	5,5	1,5	fort1	5
Finlande	0,3	5,8	0,1	119,2	18803	18,4	2,2	fort2	6

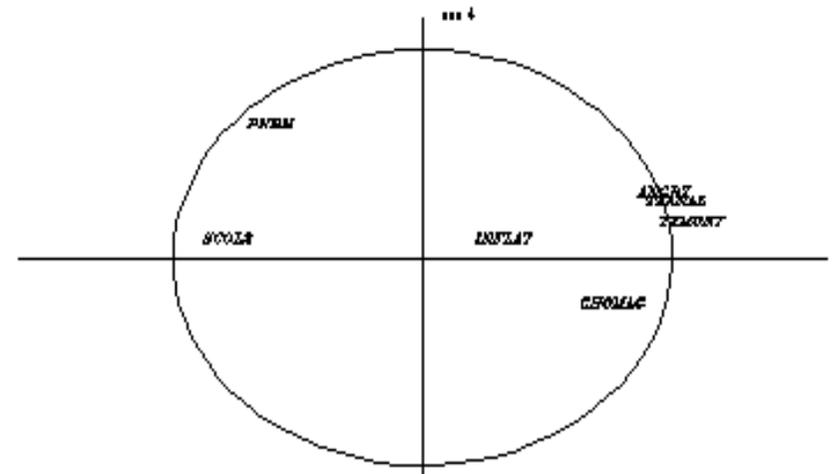
Analyse en Composantes Principales

CERCLE de CORRELATION

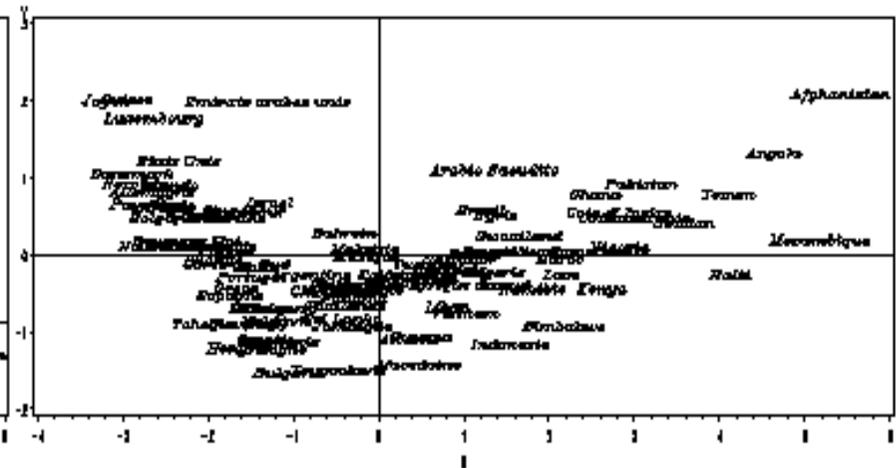
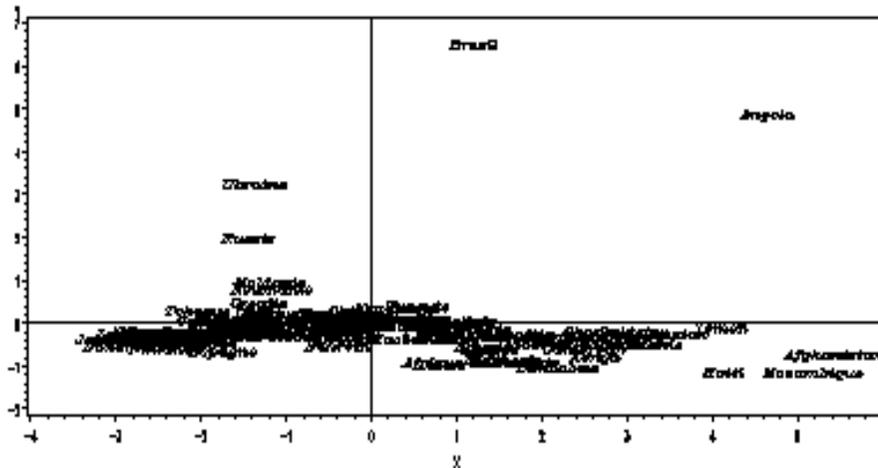


Axes 1 et 2

CERCLE de CORRELATION



Axes 1 et 4



Exemple : 96 pays classés suivant 7 variables classiques

KACP : grille 6x6 et 500 iterations

Angola Brésil			Emirats arabes uni Israël Singapour	Etats Unis Islande	Japon Luxembourg Suisse
	Comores Ghana Maroc Pakistan	Arabie Saoudite Salvador Syrie	Bahrein	Australie Canada	Allemagne Belgique Danemark Finlande France Norvège Pays Bas Suède
Afghanistan Haïti Mozambique Soudan Yemen	Cote d'Ivoire Mauritanie	Bolivie Paraguay Jordie Vietnam	Chili Malaisie Panama Philippines	Grèce	Espagne Irlande Italie Nouvelle Zélande Royaume Uni
Cameroun Nigeria	Egypte Laos Nicaragua Swaziland	Mongolie Perou Turquie	Sri Lanka	Chypre Corée du Sud Malte Portugal Tchéquie (Rep) Uruguay	Argentine
Kenya Namibie	Algérie Irak	Jamaïque	Bulgarie Croatie Hongrie Pologne Slovénie Yougoslavie		Colombie Equateur Fidji Mexique
Indonésie Zimbabwe	Afrique du sud Liban Macédoine		Moldavie Roumanie Russie Ukraine		Albanie Chine Costa Rica Guyana Thaïlande Venezuela

Classes et distances

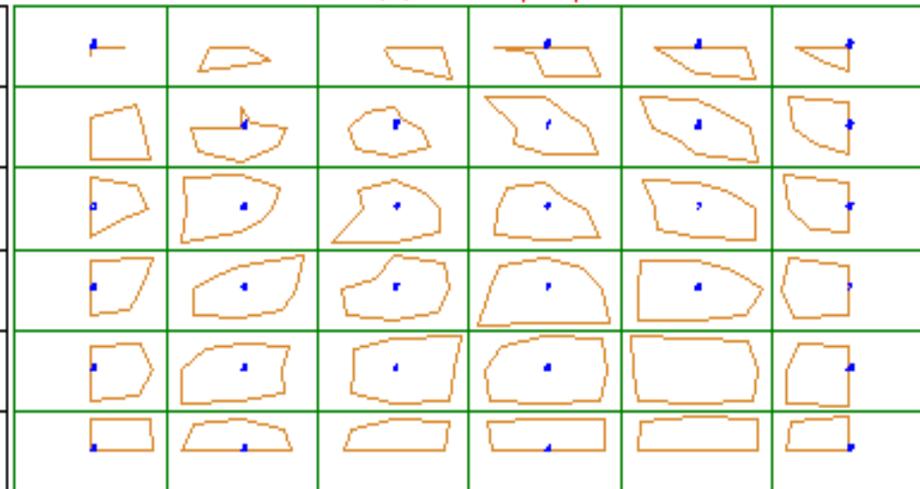
- ✓ Comme le nombre de classes est fixé a priori assez grand, il est utile de procéder à un regroupement
- ✓ On fait une classification hiérarchique sur les vecteurs codes, ce qui définit des super-classes
- ✓ On **colorie ces super-classes** (cf. classification mixte)
- ✓ On peut visualiser les distances entre les classes de Kohonen, car la disposition sur la grille donne une impression fautive d'équidistance
- ✓ Plus il y a du blanc entre deux classes (dans les 8 directions), plus la distance est grande

Classes, super-classes, distances

KACP : grille 8x8 et 500 iterations

Y900			Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900
	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900
Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900
Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900
Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900
Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900

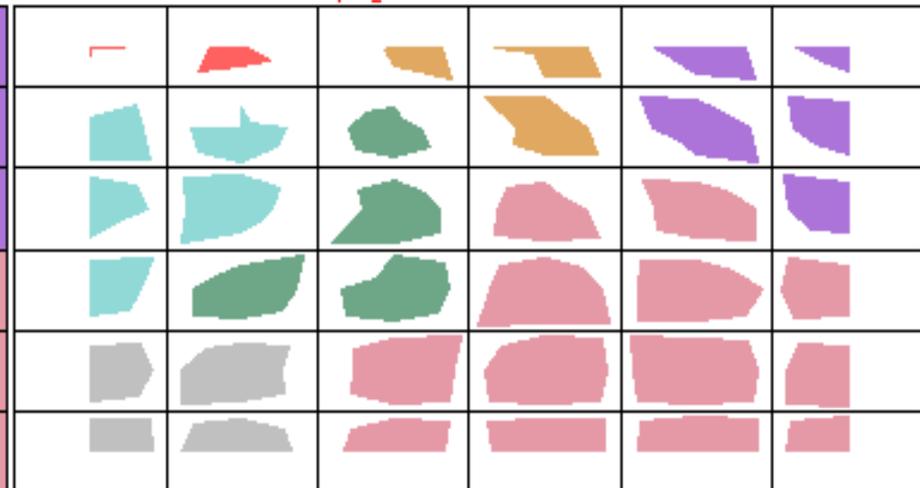
Distances (M) avec les plus proches voisins



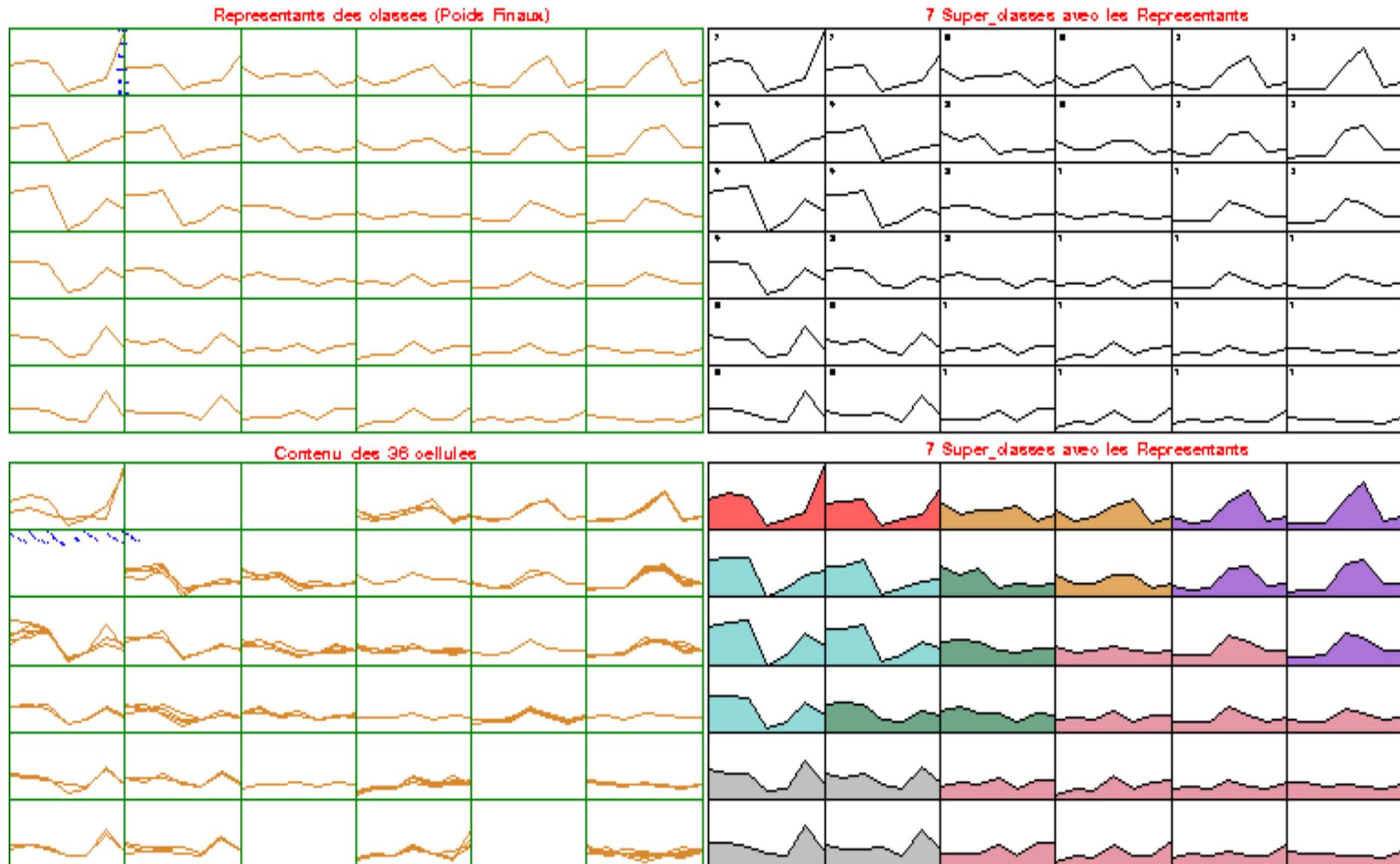
Libelles des 7 clusters

Y900			Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900
	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900
Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900
Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900
Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900
Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900	Y900 Y900 Y900

7 Super_classes avec les Proches Voisins

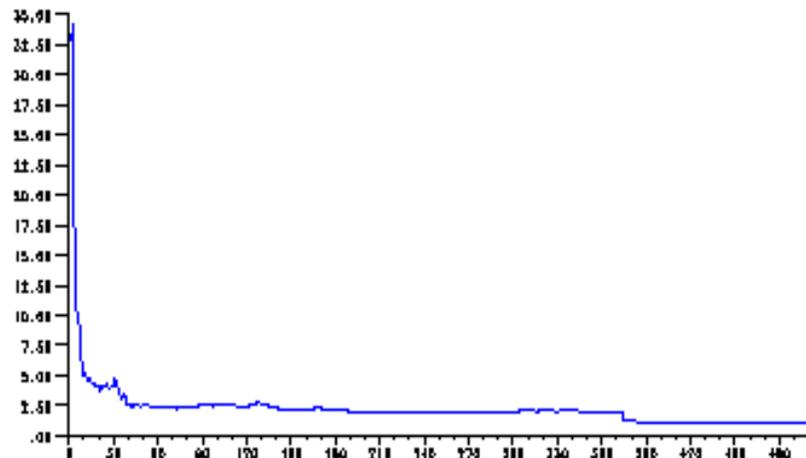


Vecteurs-codes, contenus des classes, et super-classes

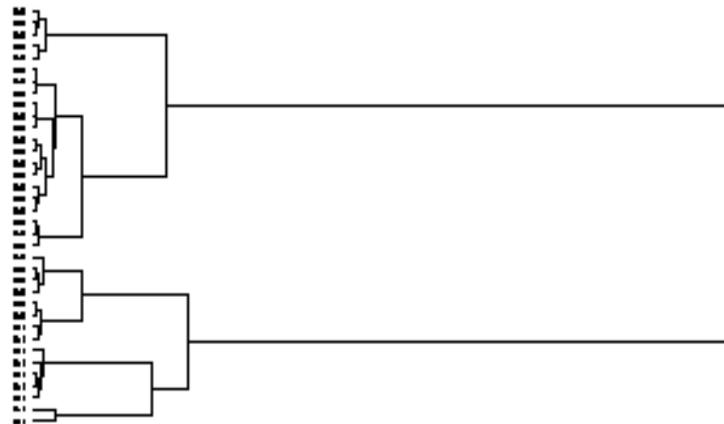


Divers

Variance intra etendue aux voisins



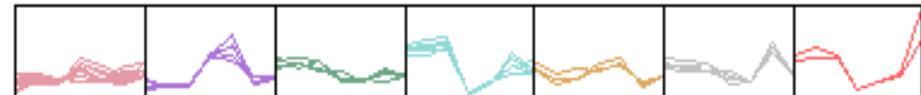
G_Dendrogramme



Representants des classes (Poids Finaux)

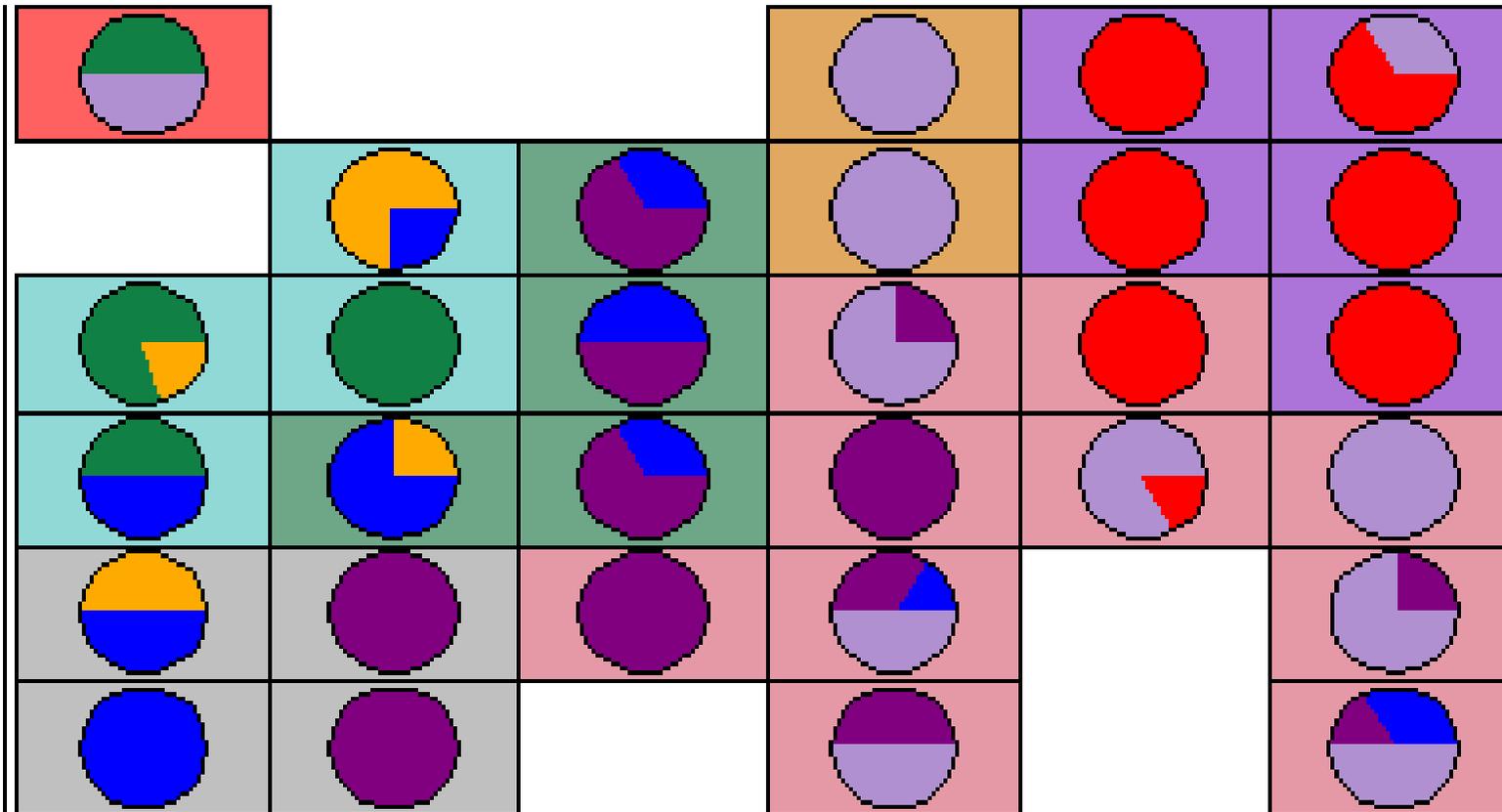


Contenu des 7 clusters



Croisement avec la variable qualitative Niveau de Développement Humain

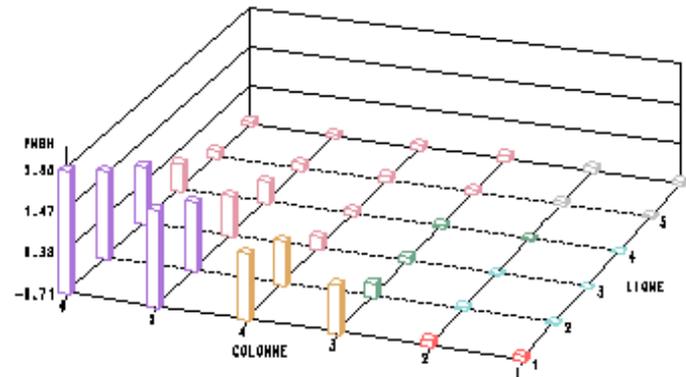
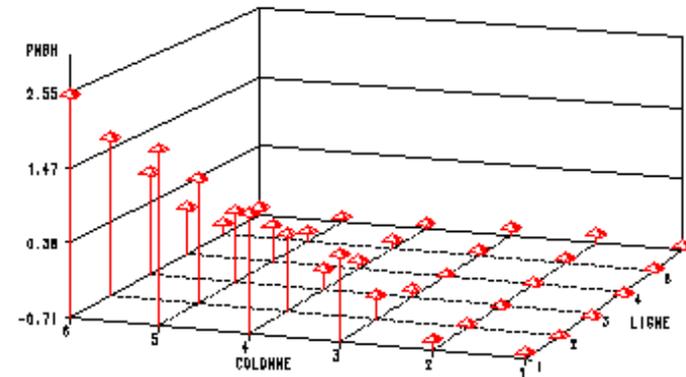
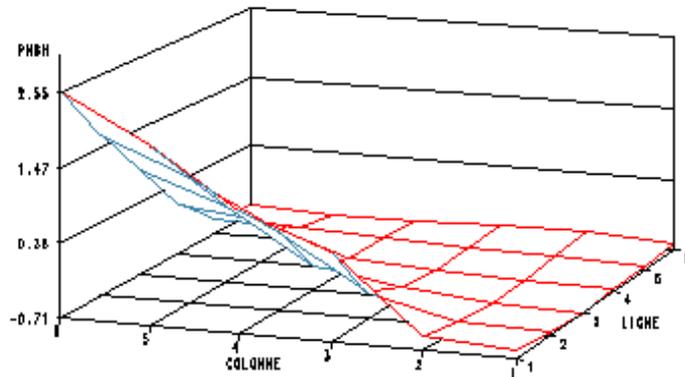
Composants de la variable : CATIDH



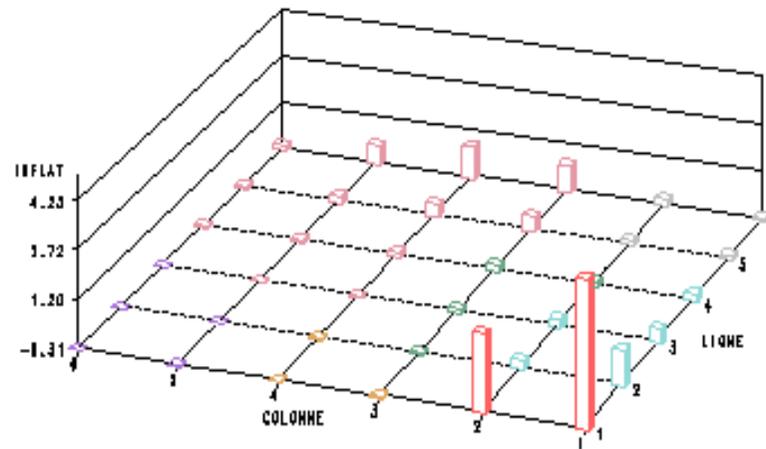
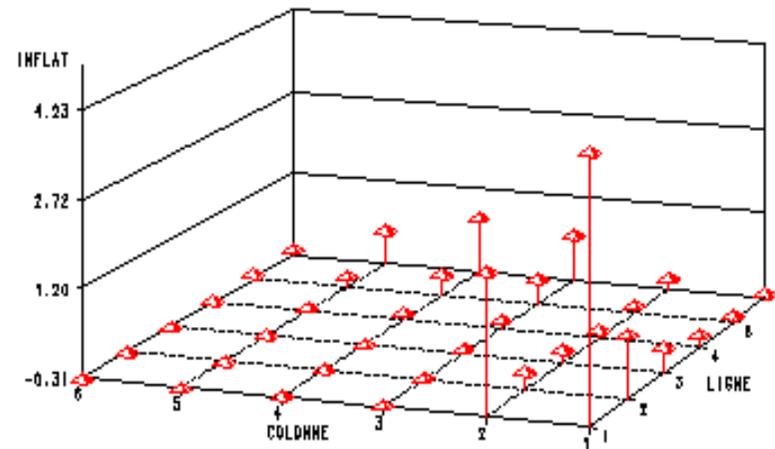
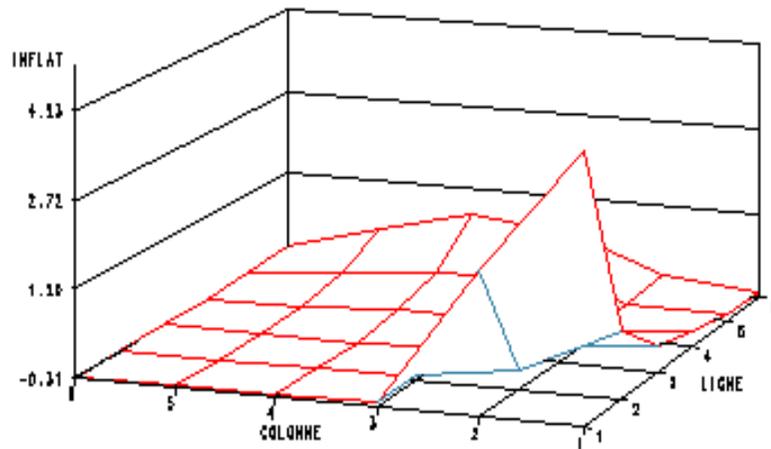
Croisement de CATIDH avec CLUSTER



PNB par habitant le long de la grille



Inflation le long de la grille



Observations avec données manquantes

PAYS	ANCRX	TXMORT	TXANAL	SCOL2	PNBH	CHOMAG	INFLAT
Andorre	2,7	6,6	0,5	21,7	13235	0,1	
Bangladesh	2,3	109	65,5	19	229		4,3
Benin	3,3	118,8	69,6	12	540		38,5
Birmanie	2,1	83,8	19,5	23	924		37
Congo	2,9	64,9	25		700	25	40,2
Coree du Nord	1,7	23,9	1		595		5,1
Cuba	0,6	9,2	3	77	580		
Dominique	0	18,3	4		2588	16	1,7
Grenade	0,4	12,5	2		2500	25	2,7
Guatemala	2,9	55,5	43	24	1029		12
Guinee	3	133,9	74		507		8
Inde	1,7	87,9	48,3	49	306		10,1
Irak	2,8	56,2	41,9	44	1165		58
Jordanie	4	33,7	17	53	1078		5
Kirghizstan	1,2	21,2	15		633	0,8	281
Koweit	0,8	16,3	21	60	14342		4
Lesotho	2,7	71,4	26,7	26	700		14
Liberia	3,3	115,8	61	5	185		11
Libye	3,4	67,9	37,2		5000		25
Liechtenstein	1,2		1		35000	1,6	5,5
Turkmenistan	2,1	43,5	14		1333		2395
Tuvalu	1,5	78,4	5		2222	12,6	106
Vanuatu	2,5	43,9	47,5	20	1212		2,3
Vatican	1,1						

Classement des observations avec données manquantes

Positions des 24 Données Supplémentaires

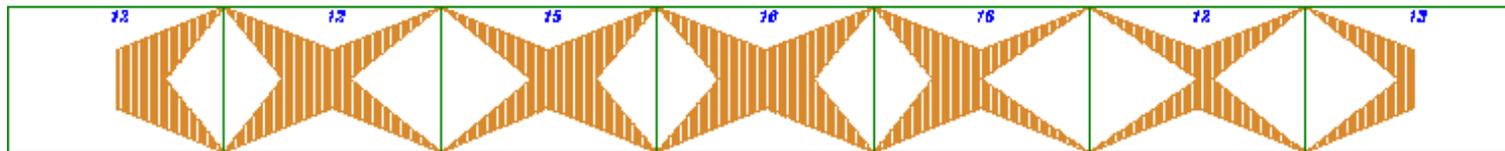
Turkmenistan			Koweït		Liechtenstein
	Vanuatu	Irak Jordanie Libye			
Benin Guinée Libéria	Bangladesh				
	Inde	Tuvalu	Vatican		
Birmanie Guatemala Lesotho	Congo		Cuba Dominique Grenade		
				Kirghizstan	Andorre Corée du Nord

KACP sur une ficelle 7

KACP : ficelle de 7 pour 500 iterations

Afghanistan Angola Comores Cote d'Ivoire Ghana Haïti Mauritanie Mozambique Nigeria	Pakistan Soudan Yemen	Algérie Cameroun Égypte Indonésie Iran Kenya Laos Maroc Namibie	Nicaragua Swaziland Zimbabwe	Afrique Albanie Arabie S. Bolivie Guyane Liban Macedoine Mongolie Paraguay	Pérou Salvador Syrie Tunisie Turquie Vietnam	Bahreïn Brésil Chine Colombie Costa R. Équateur Inde Jamaïque Malaisie	Mexique Panama Philippines Sri Lanka Thaïlande Venezuela Yougoslavie	Argentine Bulgarie Chili Corée du Sud Croatie Hongrie Malte Moldavie Pologne	Portugal Roumanie Russie Slovenie Tchéquie Ukraine Uruguay	Australie Canada Chypre Espagne Finlande Grèce Irlande Israël Italie	Nouvelle Zélande Royaume Uni Singapour	Allemagne Belgique Danemark Émirats Arabes Unis France Islande Japon Luxembourg	Norvège Pays-Bas Suède Suisse
--	-----------------------------	---	------------------------------------	--	---	--	--	--	--	--	--	--	--

Distances (M) avec les plus proches voisins



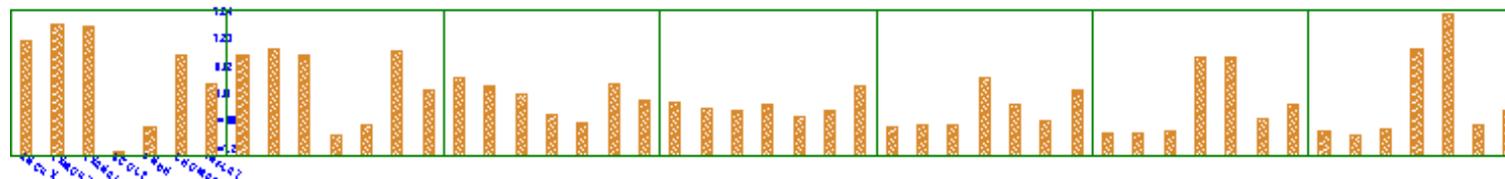
Contenu des 7 cellules



Representants des classes (Poids Finaux)



Representants des classes (Poids Finaux)



Analyse de données : introduction

Algorithme de Kohonen

Kohonen et classification : KACP

Accélération de la classification

Traitements des variables qualitatives

Conclusion

Kohonen et Classification

- ✓ Les algorithmes usuels (sans voisinage) minimisent la somme des carrés intra-classes, alors que l'algorithme de Kohonen minimise la variance intra-classes étendue
- ✓ Mais en pratique, au cours des itérations, on fait décroître le nombre de voisins jusqu'à 0 voisin. Alors l'algorithme de Kohonen est utilisé comme ***une très bonne initialisation d'un algorithme de classification usuel, qui permet d'atteindre un «bon» minimum de la variance intra-classes***

KACP pour accélérer SCL

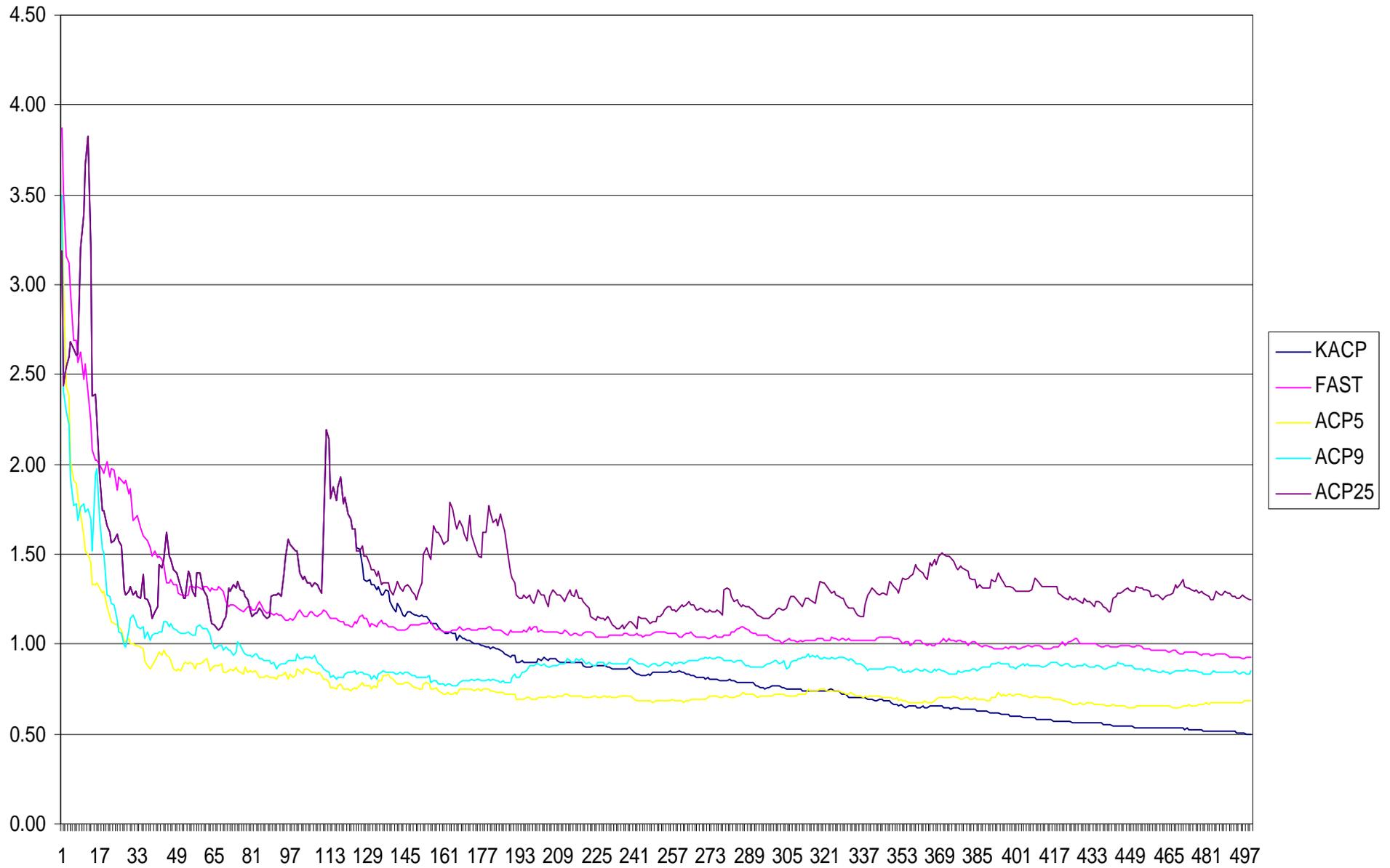
- ✓ On constate que la somme des carrés intra qu'on cherche à minimiser décroît plus vite lorsqu'on utilise un algorithme avec des voisins, que si l'on utilise le SCL (version stochastique de l'algorithme des centres mobiles)
- ✓ On réalise donc la classification de données
 - avec SCL (0 voisin)
 - avec un SOM avec 5 voisins
 - avec un SOM avec 9 voisins
 - avec un SOM avec 25 voisins
 - avec l'algorithme KACP (nb de voisins décroissant de 25 à 0 voisins, suivant la décroissance usuelle)

Les données : SAVING

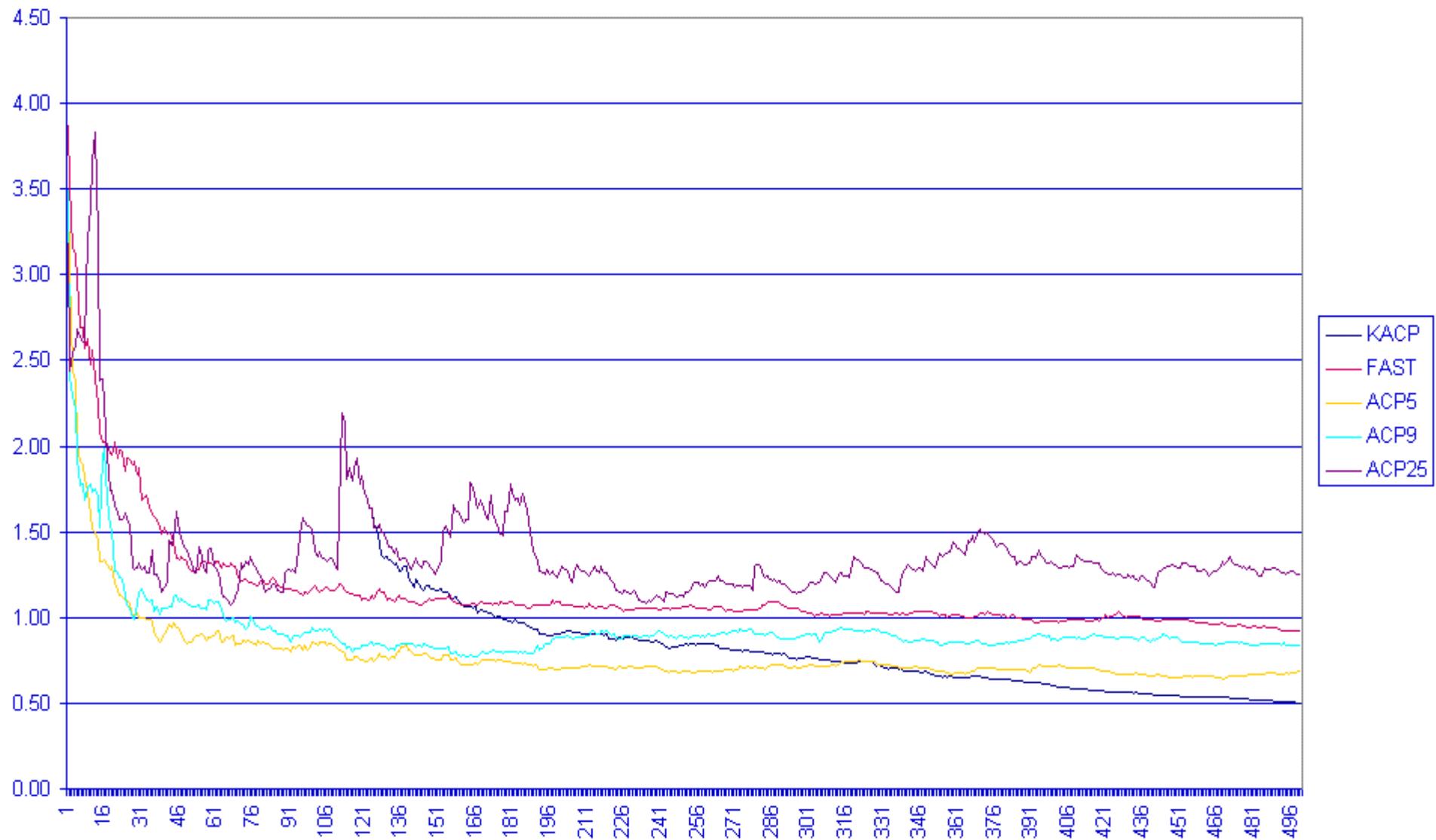
- ✓ Source : Belsey, Kuh, Welsch : Regression diagnostics, Wiley (1980)
- ✓ 42 pays, période 1960-1970
- ✓ SR : Taux moyen d'épargne par personne dans le pays (1960-1970)
- ✓ POP15 : Pourcentage moyen de population de moins de 15 ans
- ✓ POP 75 : Pourcentage moyen de population de plus de 75 ans
- ✓ DPI : Taux moyen de revenu disponible par personne
- ✓ Δ DPI : Taux moyen de croissance de DPI

Country	SR	POP15	POP75	GPI	IADPI
Australia	11.43	29.35	2.87	2329.68	2.87
Austria	12.07	23.32	4.41	1507.99	3.93
Belgium	13.17	23.80	4.43	2108.47	3.82
Bolivia	5.75	41.89	1.67	189.13	0.22
Brazil	12.88	42.19	0.83	728.47	4.56
Canada	8.79	31.72	2.85	2982.88	2.43
Chile	0.60	39.74	1.34	662.86	2.67
China (Taiwan)	11.90	44.75	0.67	289.52	6.51
Colombia	4.98	46.64	1.06	276.65	3.08
Costa Rica	10.78	47.64	1.14	471.24	2.80
Denmark	16.85	24.42	3.93	2496.53	3.99
Ecuador	3.59	46.31	1.19	287.77	2.19
Finland	11.24	27.84	2.37	1681.25	4.32
France	12.64	25.06	4.70	2213.82	4.52
Germany (F.R.)	12.55	23.31	3.35	2457.12	3.44
Greece	10.67	25.62	3.10	870.85	6.28
Guatemala	3.01	46.05	0.87	289.71	1.48
Honduras	7.70	47.32	0.58	232.44	3.19
Iceland	1.27	34.03	3.08	1900.10	1.12
India	9.00	41.31	0.96	88.94	1.54
Ireland	11.34	31.16	4.19	1139.95	2.99
Italy	14.28	24.52	3.48	1390.00	3.54
Japan	21.10	27.01	1.91	1257.28	8.21
Korea	3.98	41.74	0.91	207.68	5.81
Luxembourg	10.35	21.80	3.73	2449.39	1.57
Malta	15.48	32.54	2.47	601.05	8.12
Norway	10.25	25.95	3.67	2231.03	3.62
Netherlands	14.65	24.71	3.25	1740.70	7.66
New Zealand	10.67	32.61	3.17	1487.52	1.76
Nicaragua	7.30	45.04	1.21	325.54	2.48
Panama	4.44	43.56	1.20	568.56	3.61
Paraguay	2.02	41.18	1.05	220.56	1.03
Peru	12.70	44.19	1.28	400.06	0.67
Philippines	12.78	46.26	1.12	152.01	2.00
Portugal	12.49	28.96	2.85	579.51	7.48
South Africa	11.14	31.94	2.28	651.11	2.19
South Rhodesia	13.30	31.92	1.52	250.96	2.00
Spain	11.77	27.74	2.87	768.79	4.35
Sweden	6.86	21.44	4.54	3299.49	3.01
Switzerland	14.13	23.49	3.73	2630.96	2.70
Turkey	5.13	43.42	1.08	389.66	2.96
Tunisia	2.81	46.12	1.21	249.87	1.13

SAVING 5-5 500



SAVING 5-5 500



Analyse de données : introduction

Algorithme de Kohonen

Kohonen et classification : KACP

Accélération de la classification

Traitements des variables qualitatives

Conclusion

Analyse des relations entre modalités de variables qualitatives

Analyse d'une table de Burt (KACM)

- ✓ Classiquement, l'analyse des correspondances des modalités de plus de 2 variables qualitatives se fait par **l'analyse des correspondances multiples, qui est une analyse en composantes principales pondérée sur la table de Burt associée**. La distance considérée est la distance du χ^2 . La table de Burt est un tableau de contingence généralisé, qui croise toutes les variables qualitatives deux à deux.
- ✓ On pratique ici un algorithme de Kohonen sur cette table de Burt, avec la même pondération et la distance du χ^2 .
- ✓ Les modalités associées se retrouvent dans la même classe ou dans des classes voisines.

Exemple de table de Burt

	Q1_1	Q1_2	Q1_3	Q2_1	Q2_2	Q3_1	Q3_2	Q3_3	Q3_4
Q1_1	4	0	0	2	2	1	0	1	2
Q1_2	0	5	0	2	3	0	1	3	1
Q1_3	0	0	3	2	1	1	2	0	0
Q2_1	2	2	2	6	0	2	2	1	1
Q2_2	2	3	1	0	6	0	1	3	2
Q3_1	1	0	1	2	0	2	0	0	0
Q3_2	0	1	2	2	1	0	3	0	0
Q3_3	1	3	0	1	3	0	0	4	0
Q3_4	2	1	0	1	2	0	0	0	3

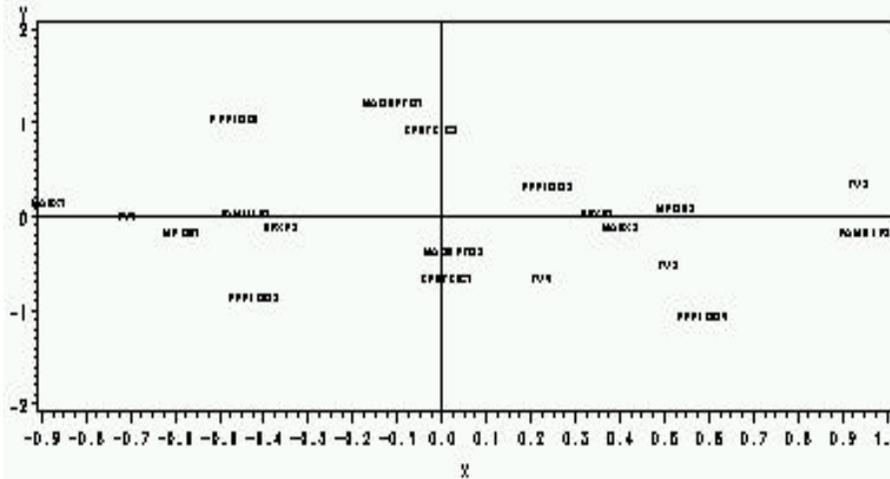
Un exemple

- ✓ Tiré de « Statistique exploratoire multidimensionnel » de Lebart, Morineau, Piron, (Dunod) 1995

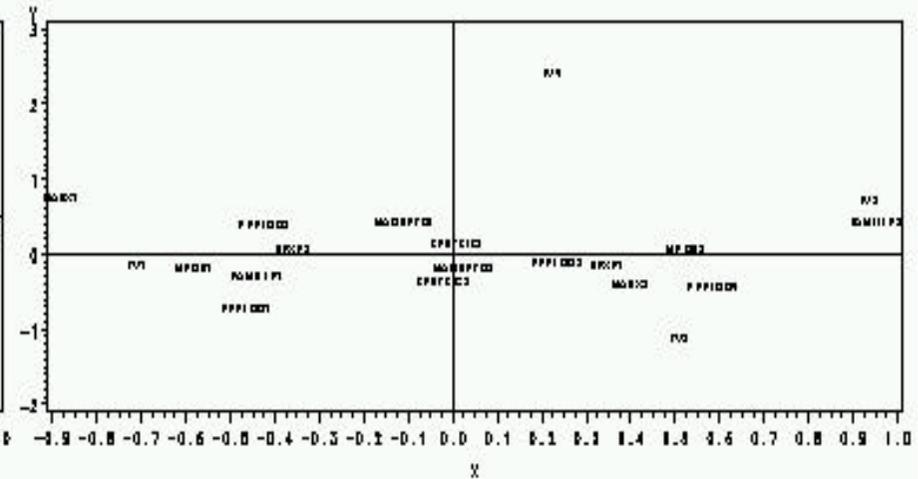
- ✓ 105 ménages, 8 questions, 20 modalités
 - La famille est l'endroit où on se sent bien : oui, non
 - Les dépenses de logement sont une charge : négligeable, sans gros problème, une lourde charge, une très lourde charge
 - Avez-vous eu récemment mal au dos : oui, non
 - Vous imposez-vous des restrictions : oui, non
 - Sexe de l'enquêté : masculin, féminin
 - avez-vous un magnétoscope : oui, non
 - Avez-vous eu récemment des maux de tête : oui, non
 - Regardez-vous la télévision : tous les jours, assez souvent, pas très souvent, jamais

Analyse des correspondances multiples

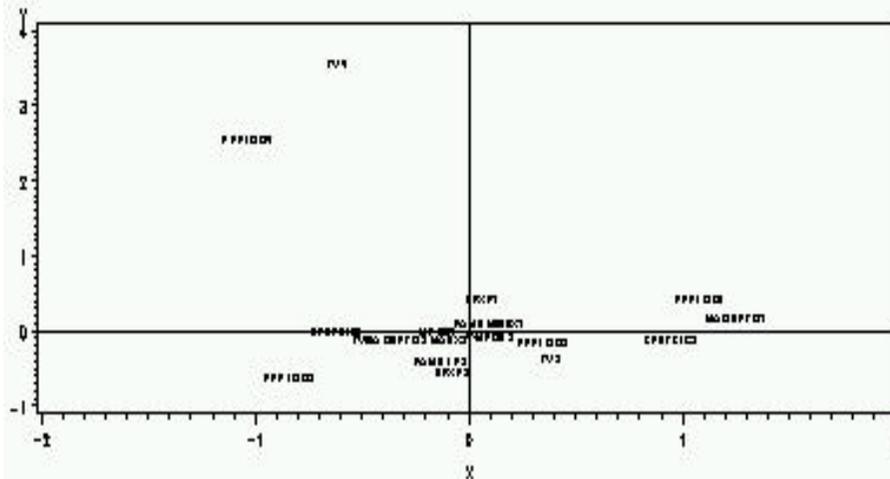
Axes 1 (0.16) et 2 (0.13)



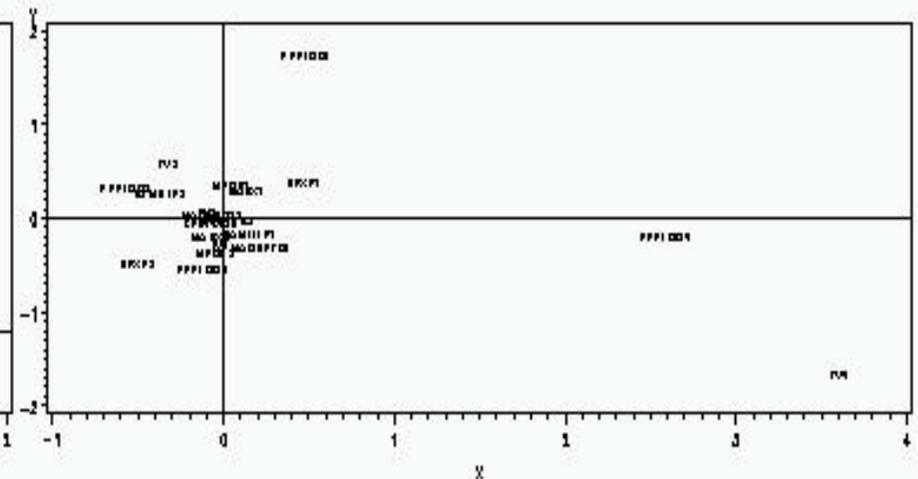
Axes 1 (0.16) et 4 (0.10)



Axes 2 (0.13) et 3 (0.11)



Axes 3 (0.11) et 5 (0.10)



Carte des modalités

KACM : grille 5x5 et 200 iterations

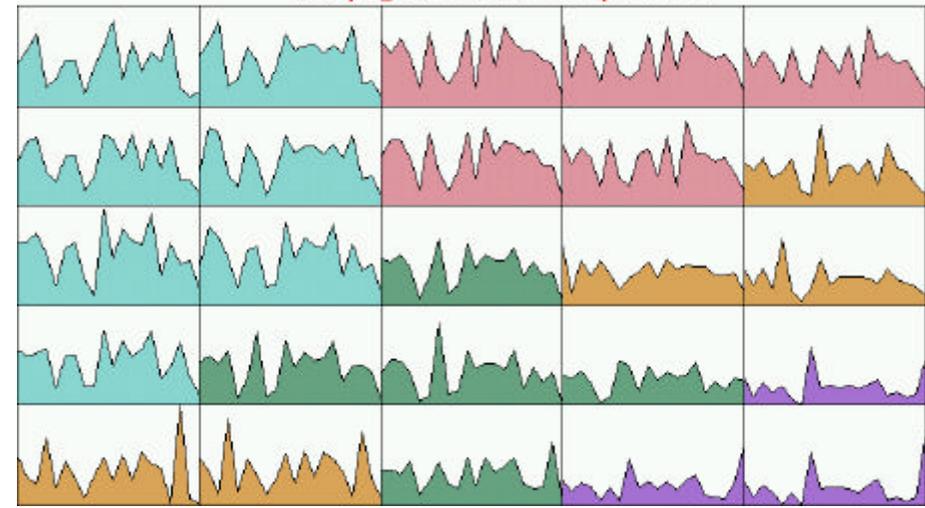
MAUX1	FAMILLE1 TV1	MAUX2	SEXE1	
MDOS1	SEXE2	DEPLOG2	MDOS2	MAGNETO1 RESTRIC2
MAGNETO2 RESTRIC1				DEPLOG1
		DEPLOG3		DEPLOG4
TV2	FAMILLE2	TV3		TV4

Super classes pour les modalités

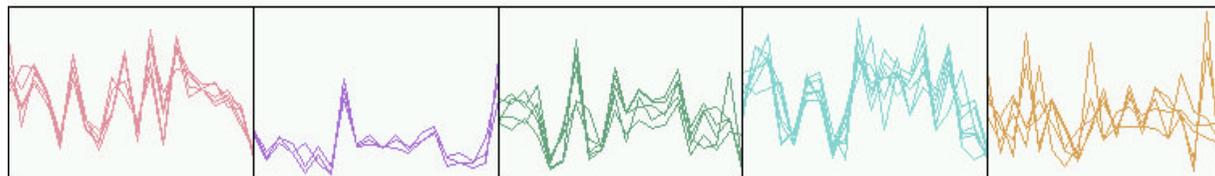
Libelles des 5 clusters

MAUX1	FAMILLE1 V1	MAUX2	SEX1	
MDOS1	SEX2	DEPLO2	MDOS2	MAGNE LO1 RESTRICT
MAGNE LO2 RESTRICT				DEPLO1
		DEPLO3		DEPLO4
V2	FAMILLE2	V3		V4

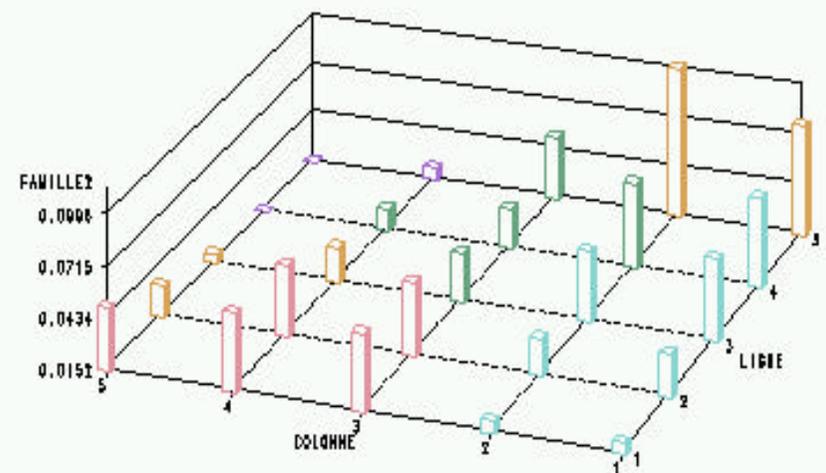
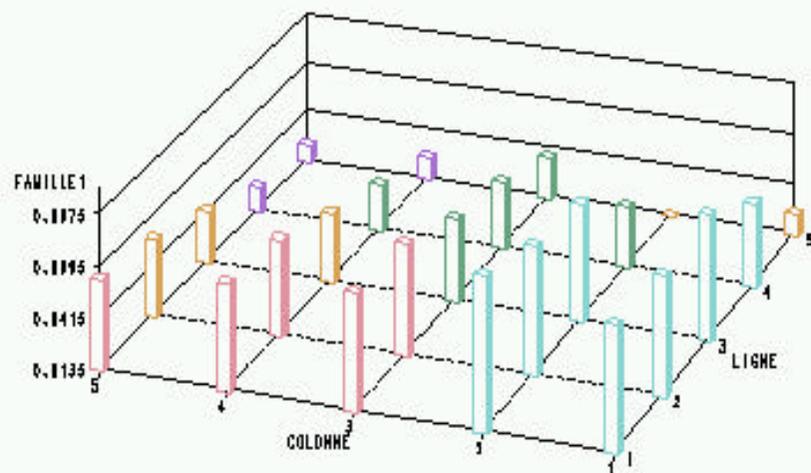
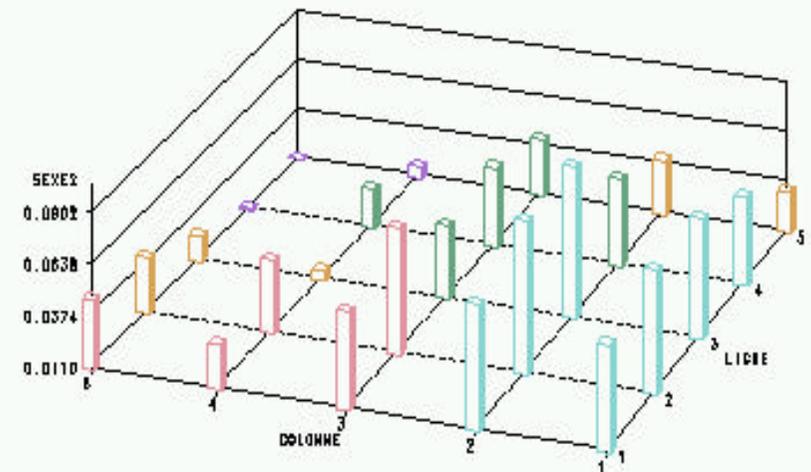
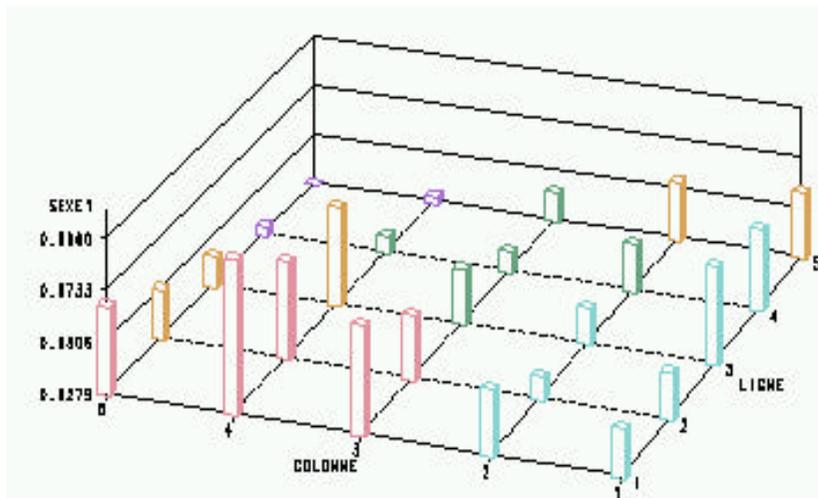
6 Super_classes avec les Representants



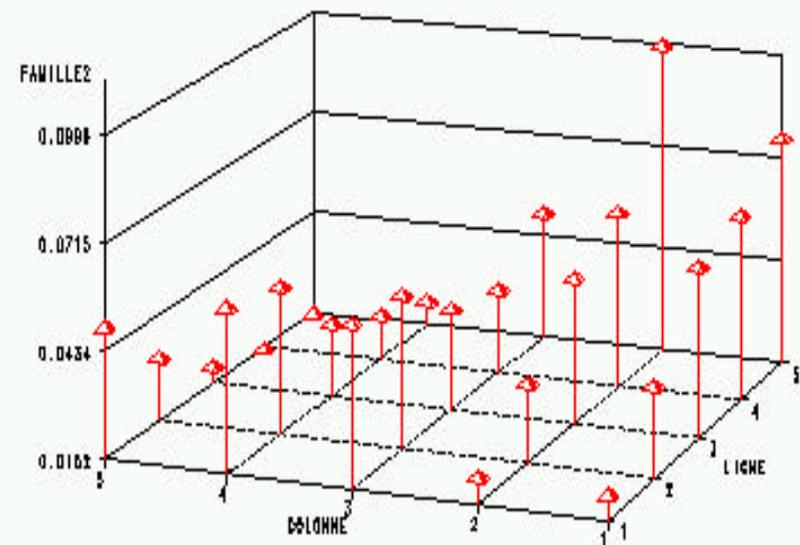
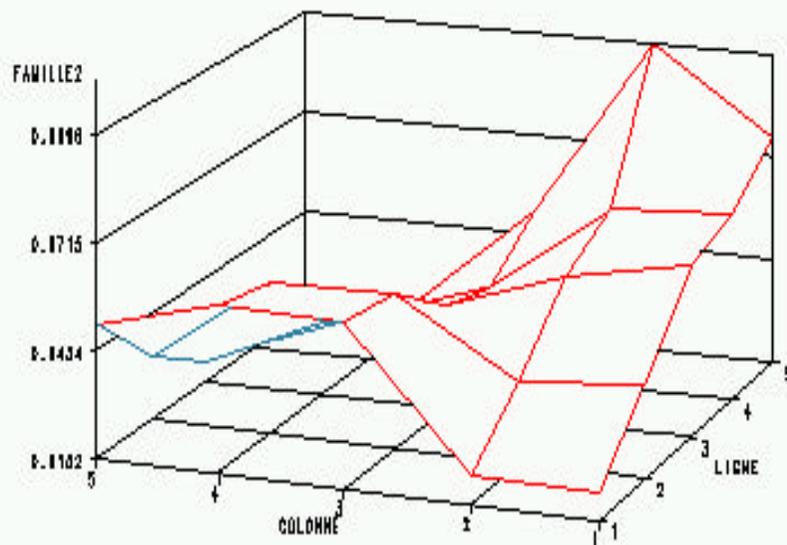
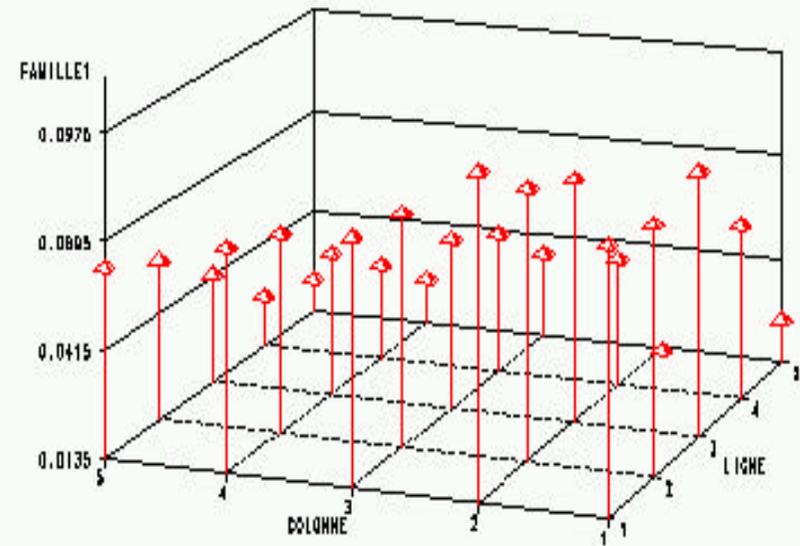
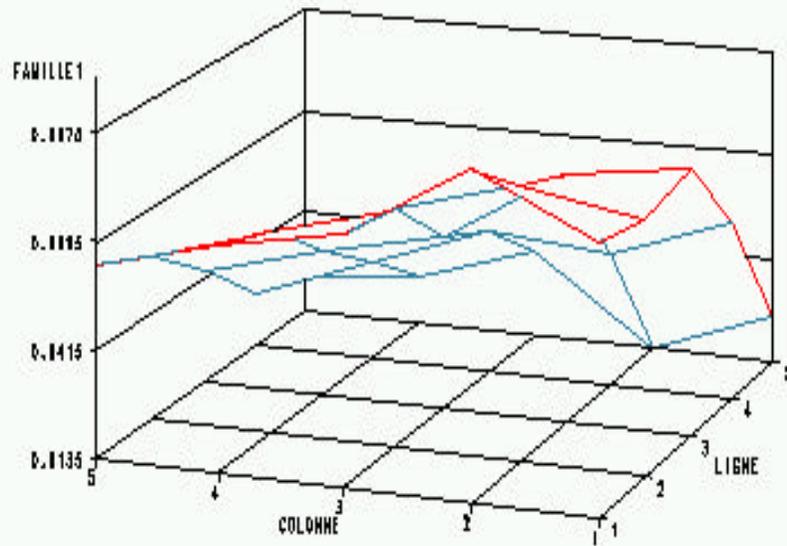
Contenu des 5 clusters



Modalités (sexe et famille) le long de la grille



Modalité famille 1 et 2



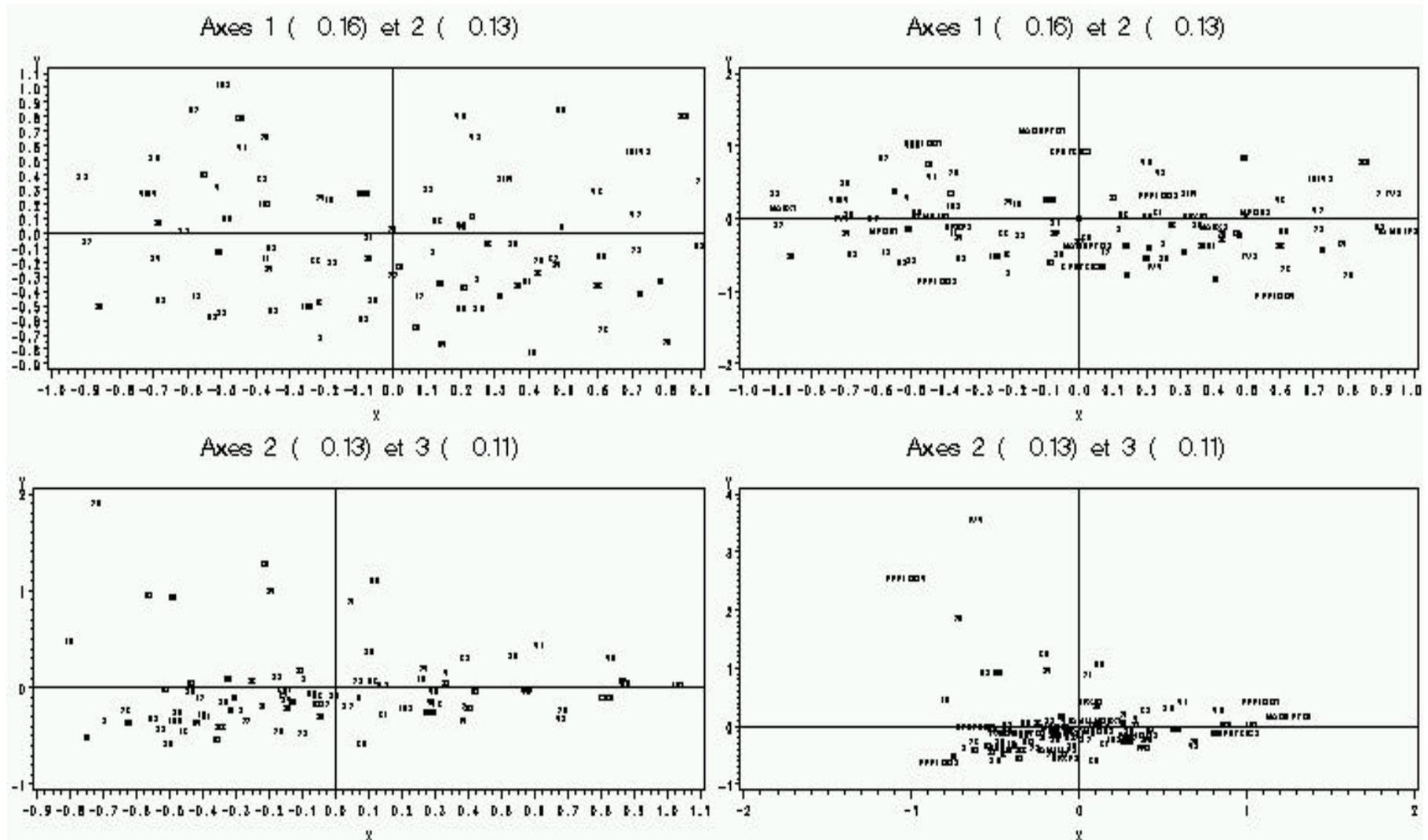
Analyse du tableau disjonctif complet : modalités et individus (KACM3)

- ✓ Si l'on souhaite représenter aussi les individus (et pas seulement les modalités), on travaille sur le tableau disjonctif complet
- ✓ Classiquement, on fait alors une analyse en composantes principales sur le tableau disjonctif complet, correctement normalisé et en utilisant la distance du χ^2 .
- ✓ La méthode KACM3 consiste alors à pratiquer un algorithme de Kohonen sur ce tableau, avec la même normalisation et la distance du χ^2 .
- ✓ Les modalités et les individus associés se retrouvent dans la même classe ou dans des classes voisines.
- ✓ La représentation graphique est malaisée (trop grand nombre de points), mais la classification obtenue est très utile.

Tableau disjonctif complet

	Q1_1	Q1_2	Q1_3	Q2_1	Q2_2	Q3_1	Q3-2	Q3_3	Q3_4
1	0	1	0	0	1	0	0	0	1
2	0	1	0	1	0	0	0	1	0
3	0	0	1	1	0	0	1	0	0
4	1	0	0	0	1	0	0	0	1
5	1	0	0	0	1	0	0	1	0
6	0	1	0	0	1	0	0	1	0
7	0	0	1	1	0	1	0	0	0
8	1	0	0	1	0	1	0	0	0
9	0	1	0	1	0	0	1	0	0
10	0	1	0	0	1	0	0	1	0
11	0	0	1	0	1	0	1	0	0
12	1	0	0	1	0	0	0	0	1

ACM : modalités et individus



Répartition des modalités et des observations

KACM3 : grille 5x5 et 800 iterations

RESTRIC1 18 61 83	FAMILLE2 10 32 64 75 87 91 94	28 38 54	TV3 2 31 88	FAMILLE1 15 27 39 82
25 43 45 52 60 74 79	DEPLOG2 3 69 84 90 98	MAUX2 1 6 19 20 73 77 81 102	SEXE2 DEPLOG3 35 78 105	MDOS2 9 49 57
MAGNETO1 37 51 67 89	8 40 58 99	MAUX1 5 11 36 47 53 72	TV1 17 22 24 29 56 100	MAGNETO2 30 42 46 104
DEPLOG4 34	DEPLOG1 95 12 97 16 23 26 41 63 85 92	66 71	MDOS1 59 78	21 44 55 62
TV4 70 80 96 103	SEXE1 65	TV2 7 50	13 48 68 93	RESTRIC2 4 14 33 86 101

Super-classes sur les modalités et observations

Libelles des 5 clusters

RESTRIC1 18 61 83	FAMILLE2 10 32 64 75 87 91 94	28 38 54	TV3 2 31 88	FAMILLE1 15 27 39 82
25 43 45 52 60 74 79	DEPLOG2 3 69 84 90 98	MAUX2 1 6 19 20 73 77 81 102	SEXE2 DEPLOG3 35 78 105	MDOS2 9 49 57
MAGNETO1 37 51 87 89	8 40 58 99	MAUX1 5 11 36 47 53 72	TV1 17 22 24 29 56 100	MAGNETO2 30 42 46 104
DEPLOG4 34	DEPLOG1 95 12 97 16 23 26 41 63 82 85 92	66 71	MDOS1 59 76	21 44 55 62
TV4 70 80 96 103	SEXE1 65	TV2 7 50	13 48 68 93	RESTRIC2 4 14 33 86 101

Valeurs tests pour KACM3

SEXE	C1	C2	C3	C4	C5	Total
1	52%	50%	78%	25%	50%	50%
2	48%	50%	22%	75%	50%	50%
FAMILLE	C1	C2	C3	C4	C5	Total
1	65%	50%	94%	85%	75%	69%
2	35%	50%	6%	15%	25%	31%
DEPLOG	C1	C2	C3	C4	C5	Total
1	4%	3%	33%	10%	25%	10%
2	48%	75%	28%	45%	50%	54%
3	48%	20%	11%	45%	0%	29%
4	0%	3%	28%	0%	25%	7%
MAGNETO	C1	C2	C3	C4	C5	Total
1	0%	25%	44%	10%	50%	21%
2	100%	75%	56%	90%	50%	79%
MAUX	C1	C2	C3	C4	C5	Total
1	30%	25%	44%	25%	75%	31%
2	70%	75%	56%	75%	25%	69%
MDOS	C1	C2	C3	C4	C5	Total
1	57%	45%	50%	40%	50%	48%
2	43%	55%	50%	60%	50%	52%
RESTRIC	C1	C2	C3	C4	C5	Total
1	52%	60%	44%	85%	75%	61%
2	48%	40%	56%	15%	25%	39%
TV	C1	C2	C3	C4	C5	Total
1	48%	45%	61%	65%	0%	50%
2	52%	25%	22%	0%	25%	26%
3	0%	30%	17%	35%	0%	21%
4	0%	0%	0%	0%	75%	3%

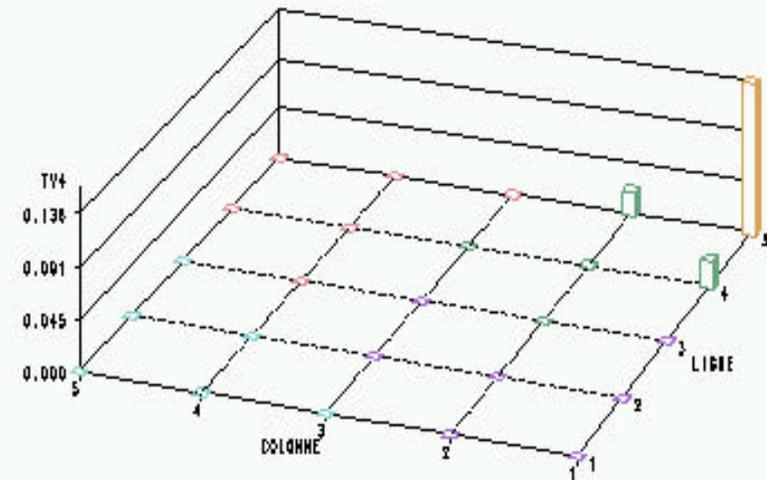
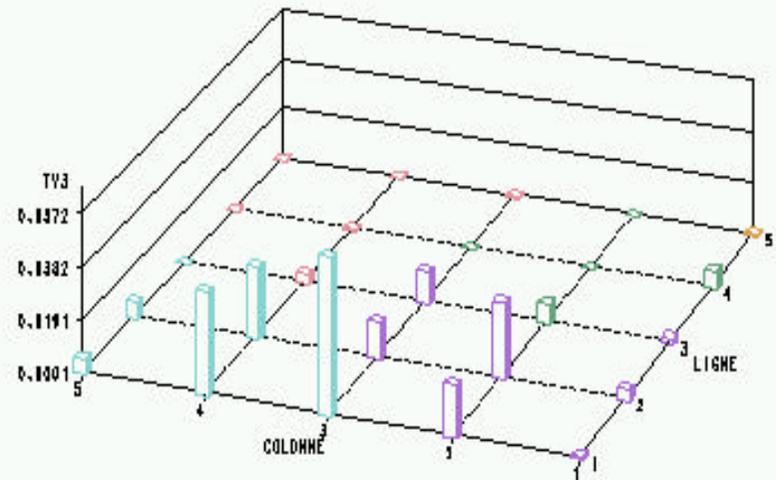
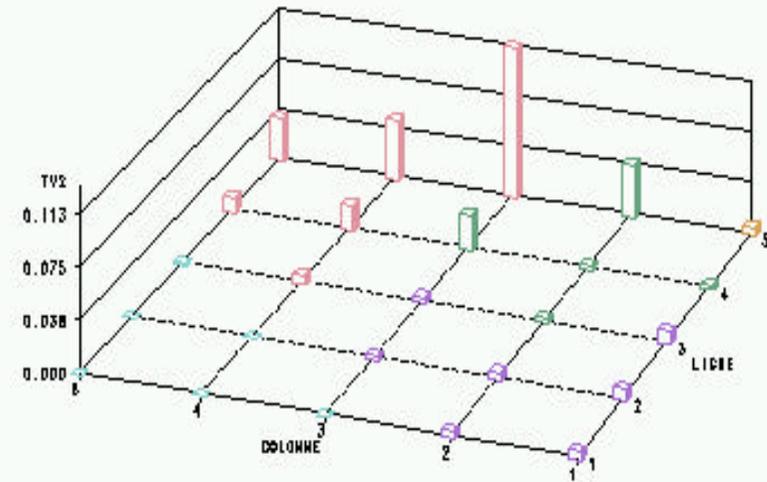
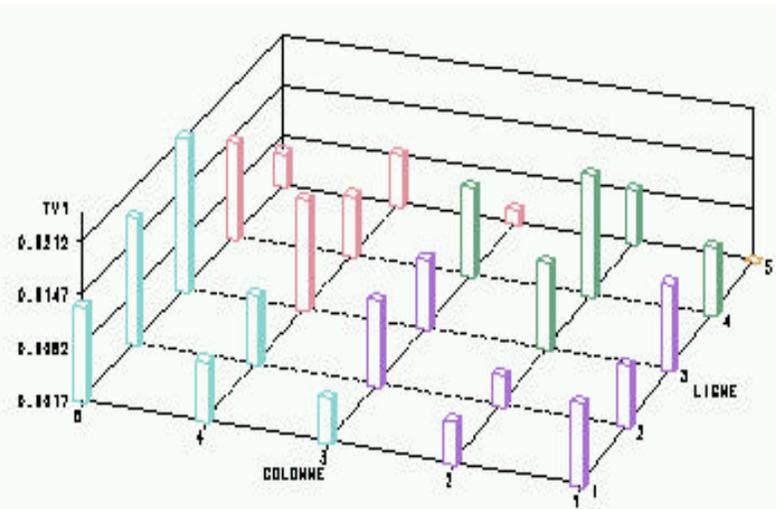
SEXE	C1	C2	C3	C4	C5
1	1.03	0.99	1.54	0.50	0.99
2	0.97	1.01	0.45	1.51	1.01
FAMILLE	C1	C2	C3	C4	C5
1	0.95	0.73	1.38	1.24	1.09
2	1.11	1.59	0.18	0.48	0.80
DEPLOG	C1	C2	C3	C4	C5
1	0.41	0.24	3.18	0.95	2.39
2	0.88	1.38	0.51	0.83	0.92
3	1.67	0.70	0.39	1.58	0.00
4	0.00	0.37	4.16	0.00	3.75
MAGNETO	C1	C2	C3	C4	C5
1	0.00	1.19	2.12	0.48	2.39
2	1.27	0.95	0.70	1.14	0.63
MAUX	C1	C2	C3	C4	C5
1	0.97	0.80	1.41	0.80	2.39
2	1.01	1.09	0.81	1.09	0.36
MDOS	C1	C2	C3	C4	C5
1	1.19	0.94	1.05	0.84	1.05
2	0.83	1.05	0.95	1.15	0.95
RESTRIC	C1	C2	C3	C4	C5
1	0.86	0.98	0.73	1.39	1.23
2	1.22	1.02	1.42	0.38	0.64
TV	C1	C2	C3	C4	C5
1	0.95	0.89	1.21	1.29	0.00
2	2.03	0.97	0.86	0.00	0.97
3	0.00	1.43	0.80	1.67	0.00
4	0.00	0.00	0.00	0.00	26.22

Distances sur la grille

5 Super_classes avec les Proches Voisins

Les modalités télévision



Analyse de données : introduction

Algorithme de Kohonen

Kohonen et classification : KACP

Accélération de la classification

Traitements des variables qualitatives

Conclusion

Conclusion

- ✓ C'est un très bon outil
 - de classification (accélération des méthodes type centres mobiles)
 - de visualisation en raison de la conservation des voisinages
 - de complément des méthodes factorielles classiques
- ✓ On peut combiner méthodes classiques et l'algorithme de Kohonen :
 - **KACP sur les coordonnées obtenues après une ACM**
 - **ACM (ou KACM) sur des variables qualitatives en y rajoutant une variable de classe obtenue par un KACP**
- ✓ On peut s'en servir en **prévision** en segmentant l'espace et en utilisant un modèle par segment (pré-traitement avant l'usage d'un perceptron ou d'un modèle auto-régressif)
- ✓ Outil de **prévision de courbes**, avec la même précision en chaque point de la courbe (au contraire des méthodes usuelles)

Conclusion

- ✓ Facilité de travail avec des **données manquantes** (cf thèse de Smaïl Ibbou) : les distances sont calculées sur les composantes présentes dans les observations
- ✓ Les données manquantes peuvent être estimées par les composantes correspondantes du vecteur code de la classe de l'observation
- ✓ Application développée par T.Kohonen : aide à la recherche de mots clés dans de grands textes (WEB)