TRAITEMENTS DE DONNEES QUALITATIVES PAR DES ALGORITHMES FONDES SUR L'ALGORITHME DE KOHONEN

Patrick Letrémy
Université de Paris 1
SAMOS-MATISSE UMR 8595
72, rue Regnault
75013 Paris, France
Tél: 01 45 83 90 67
pley@univ-paris1.fr

Résumé :

Il est bien connu que l'algorithme SOM réalise une classification des données qui peut, du fait de sa propriété de conservation de la topologie, s'interpréter comme une extension de l'Analyse en Composantes Principales. Cependant l'algorithme SOM s'applique uniquement à des données quantitatives. Dans des précédents papiers, nous avons proposé différentes méthodes basées sur l'algorithme SOM pour analyser des données d'enquête qualitatives (catégorielles). Dans ce papier, nous présentons ces méthodes de manière unifiée. La première (Kohonen Analyse des Correspondances Multiples, KACM) ne gère que les modalités, alors que les deux autres (Kohonen Analyse des Correspondances Multiples avec individus, KACM_ind et Kohonen sur le tableau DISJonctif complet, KDISJ) prennent en compte les individus et les modalités simultanément.

Mots clefs : ***Analyse des données, Données qualitatives, Cartes de Kohonen, Visualisation, Data Mining***

Abstract :

It is well known that the SOM algorithm achieves a clustering of data which can be interpreted as an extension of Principal Component Analysis, because of its topology-preserving property. But the SOM algorithm can only process real-valued data. In previous papers, we have proposed several methods based on the SOM algorithm to analyze categorical data, which is the case in survey data. In this paper, we present these methods in a unified manner. The first one (Kohonen Multiple Correspondence Analysis, KMCA) deals only with the modalities, while the two others (Kohonen Multiple Correspondence Analysis with individuals, KMCA_ind, Kohonen algorithm on DISJonctive table, KDISJ) can take into account the individuals, and the modalities simultaneously.

Key-Words : Data Analysis, Categorical Data, Kohonen Maps, Data Visualization, Data Mining.

Ce papier est très largement inspiré de l'article de Cottrell, Ibbou, Letrémy (2004) à paraître dans la revue Neural Network., et de Cottrell, M., Ibbou, S., Letrémy, P., Rousset, P. (2003).

Remerciements :

Je remercie chaleureusement Marie Cottrell pour notre Amicale et Fructueuse Collaboration.

TRAITEMENTS DE DONNEES QUALITATIVES PAR DES ALGORITHMES FONDES SUR L'ALGORITHME DE KOHONEN

Résumé :

Il est bien connu que l'algorithme SOM réalise une classification des données qui peut, du fait de sa propriété de conservation de la topologie, s'interpréter comme une extension de l'Analyse en Composantes Principales. Cependant l'algorithme SOM s'applique uniquement à des données quantitatives. Dans des précédents papiers, nous avons proposé différentes méthodes basées sur l'algorithme SOM pour analyser des données d'enquête qualitatives (catégorielles). Dans ce papier, nous présentons ces méthodes de manière unifiée. La première (Kohonen Analyse des Correspondances Multiples, KACM) ne gère que les modalités, alors que les deux autres (Kohonen Analyse des Correspondances Multiples avec individus, KACM_ind et Kohonen sur le tableau DISJonctif complet, KDISJ) prennent en compte les individus et les modalités simultanément.

Mots clefs : ***Analyse des données, Données qualitatives, Cartes de Kohonen, Visualisation, Data Mining***

Abstract :

It is well known that the SOM algorithm achieves a clustering of data which can be interpreted as an extension of Principal Component Analysis, because of its topology-preserving property. But the SOM algorithm can only process real-valued data. In previous papers, we have proposed several methods based on the SOM algorithm to analyze categorical data, which is the case in survey data. In this paper, we present these methods in a unified manner. The first one (Kohonen Multiple Correspondence Analysis, KMCA) deals only with the modalities, while the two others (Kohonen Multiple Correspondence Analysis with individuals, KMCA_ind, Kohonen algorithm on DISJonctive table, KDISJ) can take into account the individuals, and the modalities simultaneously.

Key-Words : Data Analysis, Categorical Data, Kohonen Maps, Data Visualization, Data Mining.

TRAITEMENTS DE DONNEES QUALITATIVES PAR DES ALGORITHMES FONDES SUR L'ALGORITHME DE KOHONEN

1. Introduction à l'analyse de données

1.1 Les méthodes classiques

Pour étudier, résumer, représenter des données multidimensionnelles comprenant à la fois des variables quantitatives (à valeurs continues réelles) et qualitatives (discrètes, ordinales ou nominales), les praticiens ont à leur disposition de très nombreuses méthodes performantes, éprouvées et déjà implantées dans la plupart des logiciels statistiques. L'analyse de données consiste à construire des représentations simplifiées de données brutes,

pour mettre en évidence les relations, les dominantes, la structure interne du *nuage* des observations. On peut distinguer deux grands groupes de techniques classiques : les *méthodes factorielles* et les *méthodes de classification*.

Les *méthodes factorielles* sont essentiellement linéaires ; elles consistent à chercher des sous-espaces vectoriels, des changements de repères, permettant de réduire les dimensions tout en perdant le moins d'information possible. Les plus connues sont *l'Analyse en Composantes Principales* (ACP) qui permet de projeter des données quantitatives sur les axes les plus significatifs et *l'Analyse des Correspondances* qui permet d'analyser les relations entre les différentes modalités de variables qualitatives croisées (Analyse Factorielle des Correspondances, AFC, pour deux variables, Analyse des Correspondances Multiples, ACM, pour plus de deux variables).

Les *méthodes de classification* sont très nombreuses et diverses. Elles permettent de grouper et de ranger les observations. Les plus utilisées sont la *Classification Hiérarchique* où le nombre de classes n'est pas fixé a priori et la *Méthode des Centres Mobiles* où l'on cherche à regrouper les données en un certain nombre de classes. La méthode des Centres Mobiles a été étendue aux méthodes dites de *nuées dynamiques.*

On trouvera par exemple dans Lebart et coll. (1995) ou Saporta (1990) une présentation de ces méthodes avec de nombreux exemples. On peut aussi se reporter à des références de base comme Lebart et coll. (1984), Benzécri (1973).

1.2 Les méthodes neuronales

Plus récemment, depuis les années 80, de nouvelles méthodes sont apparues, connues sous le nom de *méthodes neuronales*. Elles proviennent de travaux pluridisciplinaires où se sont retrouvés des biologistes, des physiciens, des informaticiens, des théoriciens du signal, des cogniticiens et, plus récemment encore, des mathématiciens et notamment des statisticiens. Le petit ouvrage de Blayo et Verleysen (1996) dans la collection Que Sais-je ? est une excellente introduction à ce domaine.

Outre le fait qu'elles sont partiellement issues d'une inspiration biologique ou cognitive, ces méthodes ont rencontré rapidement un certain succès en particulier à cause de leur caractère de « boîte noire », d'outil à tout faire, ayant de très nombreux domaines d'applications. Une fois dépassés un certain excès d'enthousiasme et des difficultés de mise en œuvre, les chercheurs et utilisateurs disposent maintenant d'un arsenal de techniques alternatives, non linéaires en général et algorithmiques. On pourra consulter par exemple l'ouvrage de Ripley (1996) qui intègre les techniques neuronales parmi les méthodes statistiques. En particulier, les statisticiens commencent à intégrer ces méthodes parmi l'ensemble de leurs outils. Voir à ce sujet les nouveaux modules neuronaux des grands logiciels de Statistique (SAS, SYLAB, STATLAB, S+) et des logiciels de calcul (MATLAB, R, GAUSS, etc.).

Le point important est que lorsque les méthodes statistiques linéaires standard ne sont pas appropriées, du fait de la structure intrinsèque des observations, les modèles neuronaux qui sont fortement non linéaires s'avèrent très utiles.

Le plus connu des modèles neuronaux reste sans conteste le modèle du *Perceptron Multicouches* (voir par exemple Rumelhart et McClelland., 1986), mais le fait qu'il nécessite un apprentissage supervisé (supposant une connaissance *a priori)* le rend inapte pour

l'interprétation, la représentation et la visualisation de données pour lesquelles il n'y a aucune connaissance *a priori.*

Ainsi en analyse des données, les méthodes non supervisées sont très attractives et en particulier l'algorithme de Kohonen est de nos jours largement utilisé dans ce contexte (Kohonen, 1984, 1993, 1995, 1997). Il réussit la double tâche de « projection » et de classification.
Rappelons la définition de cet algorithme, aussi appelé algorithme SOM (Self-Organizing Map). Dans sa forme originelle, il traite des données quantitatives à valeurs continues réelles, pour lesquelles chaque observation est décrite par un vecteur réel. Par exemple les variables quantitatives peuvent être des ratios, des quantités, des mesures, des indices, codés par des nombres réels. Pour le moment, nous ne considérons pas les variables qualitatives qui peuvent être présentes dans la base de données.

Dans ce cas (uniquement pour des variables quantitatives), nous considérons un ensemble de N observations, dans lequel chaque individu est décrit par p variables quantitatives à valeurs réelles. L'outil principal est un réseau de Kohonen, généralement une grille bidimensionnelle de n par n unités, ou une ficelle unidimensionnelle de n unités. Les données sont rangées dans une table X de N lignes et de p colonnes. Les lignes de la table X sont les entrées de l'algorithme SOM. Après apprentissage, chaque unité u est représentée dans l'espace R^p par son vecteur poids C_u (ou *vecteur code*). Puis chaque observation est classée selon la méthode du plus proche voisin: l'observation i appartient à la classe u si et seulement si le vecteur code C_u est le plus proche parmi tous les vecteurs codes. La distance dans R^p est en général la distance euclidienne, mais selon l'application on peut utiliser d'autres distances.

Comparée à n'importe quelle autre méthode de classification, la principale caractéristique de la classification de Kohonen est la conservation de la topologie : Après apprentissage, des observations « proches » sont associées à la même classe ou à des classes « proches » selon la définition du voisinage dans le réseau de Kohonen.

Il y a un très grand nombre d'applications de cet algorithme à des données quantitatives à valeurs réelles, voir par exemple les pages WEB à http://www.cis.hut.fi/ ou http://samos.univ-paris1.fr/

Toutes ces études montrent que l'algorithme SOM possède de nombreuses propriétés : la représentation en grille ou en ficelle est facile à interpréter, la conservation de la topologie fournit un « ordre » aux différentes classes, il est possible d'utiliser des données avec valeurs manquantes et l'algorithme de classification est rapide et efficace (De Bodt et al., 2003).

1.3 Carte de Kohonen versus l'ACP

La carte de Kohonen construite à partir des lignes de la table X peut être comparée aux projections linéaires réalisées par l'Analyse en Composantes Principales (ACP) sur les axes principaux successifs. Voir par exemple Blayo et Desmartines (1991, 1992) ou Kaski (1997). Pour trouver ces axes, il faut calculer les valeurs propres et les vecteurs propres de la matrice $X'X$ ou X' est la transposée de la matrice X. Chaque valeur propre est égale à la part de l'inertie totale représentée sur l'axe correspondant. Les axes sont ordonnés selon les valeurs propres décroissantes. Ainsi les deux premiers axes correspondent aux deux plus grandes valeurs propres, et fournissent la meilleure projection. Cependant, il est souvent nécessaire de

prendre en compte plusieurs projections bidimensionnelles de l'ACP pour avoir une bonne représentation des données, alors qu'il y a qu'une seule carte de Kohonen.

Insistons sur le point suivant : si *X* est la matrice des données centrées, l'ACP est réalisée via la diagonalisation de la matrice *X'X*, alors que la carte de Kohonen est construite avec les lignes de la matrice *X*.

2. Commentaires préliminaires à propos des variables qualitatives

2.1 Les variables qualitatives

Dans les applications réelles, les individus peuvent aussi être décrits par des variables de nature qualitative (ou catégorielle). C'est par exemple le cas pour les données d'enquête, où les gens ont à répondre à des questions ayant un nombre fini de modalités de réponse (i.e. sexe, csp, niveau de revenu, type d'emploi, lieu d'habitation, niveau d'éducation, etc.).

Attention, la plupart du temps, les variables qualitatives ne peuvent pas être utilisées telles quelles, même lorsque les modalités sont codées par des nombres. S'il n'existe pas de relation d'ordre sur les codes (1 pour yeux bleus, 2 pour yeux marrons, etc.), cela n'a aucun sens de les utiliser comme des variables numériques pour faire un apprentissage de Kohonen. Même si les codes correspondent à une progression croissante ou décroissante, cela n'aurait un sens que si une échelle linéaire était utilisée (la modalité 2 correspondant à la moitié de la progression entre la modalité 1 et 3). *Ainsi les données qualitatives nécessitent un traitement spécifique*. Quand la base des données inclut des variables quantitatives et qualitatives, la première idée est de faire une classification sur les variables quantitatives, et puis de croiser cette classification avec les autres variables qualitatives.

2.2 Croisement d'une classification avec des variables qualitatives

Supposons qu'à l'aide de l'algorithme de Kohonen, les individus aient été regroupés en classes à partir des variables quantitatives qui les décrivent.
Pour interpréter les classes selon des variables qualitatives non utilisées dans l'algorithme de classement de Kohonen, il peut être intéressant d'étudier la répartition de leurs modalités dans chaque classe. Après avoir calculé des statistiques élémentaires dans chaque classe, on peut dessiner à l'intérieur de chaque cellule un camembert montrant comment sont réparties les modalités de chacune des variables qualitatives, comme le montre la Fig. 1.

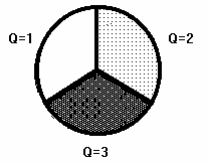

Fig 1 : Camembert, pour la variable qualitative Q possédant trois modalités équi-réparties.

2.3 Exemple I : les 96 pays en 1996 (POP_96)

On prend un exemple classique . Les données de POP _96 contiennent 7 ratios mesurés en 1996 à partir de la situation macroéconomique de 96 pays : croissance annuelle de la population, taux de mortalité infantile, taux d'analphabétisme, indice de fréquentation scolaire du second degré, PNB par tête, taux de chômage et taux d'inflation. Ces 96 pays décrits par ces 7 variables réelles sont dans un premier temps classés sur une carte de Kohonen (grille de 6x6 unités).

Le choix initial du nombre de classes est arbitraire et il n'existe pas de méthode sûre pour choisir la taille de la grille. Pour exploiter les caractéristiques stochastiques de l'algorithme SOM, et obtenir une bonne classification et une bonne organisation, il est préférable de travailler sur des grandes cartes. Mais on peut penser que le nombre significatif de classes sera souvent plus petit que la taille de la grille (nxn). D'un autre côté, il n'est ni facile ni utile d'interpréter et de décrire un trop grand nombre de classes. Aussi proposons-nous (Cottrell et al., 1997) de réduire le nombre de classes en utilisant une Classification Hiérarchique Ascendante sur les vecteurs codes avec la distance de Ward (Lance et Williams, 1967, Anderberg, 1973).

De cette manière, nous définissons deux classifications emboîtées, ce qui nous permet de distinguer les classes de Kohonen (ou « micro-classes ») et les « macro-classes » qui regroupent certaines « micro-classes ». Pour rendre claire cette classification à deux niveaux, nous associons à chaque macro-classe une couleur ou un niveau de gris.
L'avantage de cette double classification est qu'elle permet d'analyser les données à un niveau « macro » qui met en évidence les caractéristiques générales et à un niveau « micro » qui permet de déterminer les caractéristiques de phénomènes plus précis et en particulier des chemins permettant d'aller d'une classe à une autre.
Dans les applications que nous avons traitées, les « macro-classes » créent toujours des surfaces d'un seul tenant dans la grille. Cette remarque est très intéressante car elle confirme les propriétés topologiques des cartes de Kohonen.
Dans la Fig. 2, on trouve une représentation des « micro-classes » regroupées afin de constituer sept « macro-classes ». La macro-classe dans le coin supérieur gauche regroupe les pays de l'OCDE riches et très développés ; les pays très pauvres apparaissent sur la droite ; les pays anciennement socialistes ne sont pas très loin des pays les plus riches ; les pays très inflationnistes sont en haut au milieu etc…

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

Fig. 2 : Les 36 classes de Kohonen, regroupées en 7 macro-classes, 600 itérations

Afin de décrire plus précisément les pays, nous pouvons par exemple considérer une variable qualitative supplémentaire, bien connue des économistes. L'IDH (Indicateur de Développement Humain) prend en compte de nombreuses variables qui qualifient le mode de vie, la culture, la sécurité, le nombre de médecins, de théâtres. La variable qualitative est ainsi l'indice IDH, avec 3 niveaux : faible (jaune), moyen (vert) et élevé (bleu). Sur la Fig. 3, on voit la répartition de cette variable qualitative supplémentaire sur la carte de Kohonen.

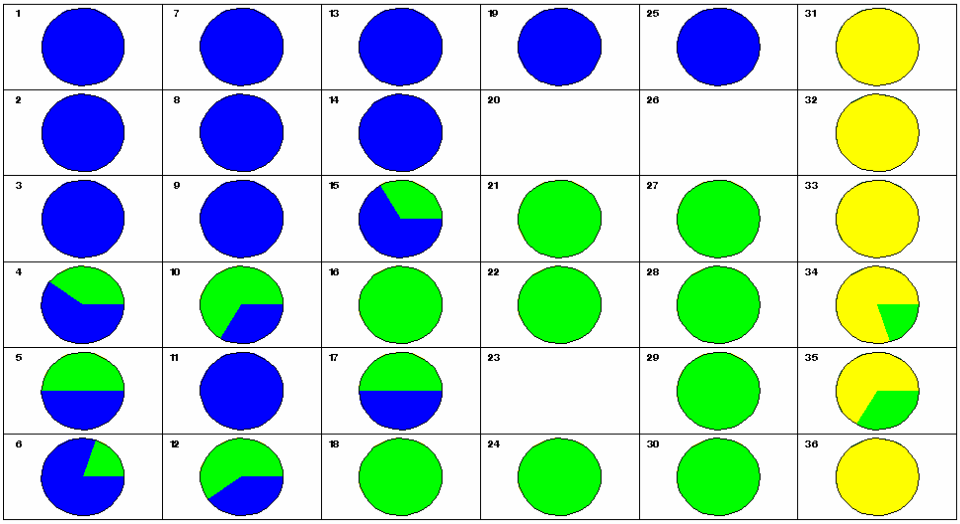

Fig. 3: Dans chaque cellule, la répartition de la variable IDH est représentée. On observe que les unités 1, 7, 13, 19, 25, 2, 8, 14, 3, 9, 15, 4, 5, 11, 17, 6 contiennent principalement des pays à fort niveau d'IDH. Les unités 31, 32, 33, 34, 35, 36 correspondant au plus faible niveau. Ceci est en complète cohérence avec la précédente classification (Fig. 2).

Dans les paragraphes 3 et 4, on considère une base de données qualitatives (catégorielles) et la méthode classique pour la traiter. Dans le paragraphe 5, on définit l'algorithme KACM, dérivé de l'algorithme de Kohonen. Il permet une première classification facile à visualiser, une bonne représentation des modalités, cette classification pouvant être regroupée en macro-classes. KACM est appliqué à des données réelles dans le paragraphe 6. Dans les paragraphes 7 et 8, on définit d'autres algorithmes qui prennent en compte les individus et les modalités. Le même exemple sera utilisé dans tous ces paragraphes. Les paragraphes 9 et 10 présentent deux autres exemples ; le premier est un exemple jouet, le second correspond à des données réelles. Le paragraphe 11 est consacré à la conclusion.

3. Base de données constituée uniquement de variables qualitatives

A partir de maintenant, toutes les observations seront décrites par K variables qualitatives, chacune ayant un certain nombre de modalités, comme dans une enquête.

Définissons les données et introduisons les notations de base. Soient un ensemble de N individus et K variables ou questions. Chaque question k ($1 \le k \le K$) possède m_k modalités (réponses, niveaux). Les individus répondent à chaque question k en choisissant seulement une modalité parmi les m_k modalités. Par exemple, si on suppose que $K = 3$ et $m_1 = 3$, $m_2 = 2$ et $m_3 = 3$, alors une réponse d'individu pourrait être (0,1,0|0,1|1,0,0), où 1 correspond à la modalité choisie pour chaque question.

Donc, si $M = \sum_{k=1}^{K} m_k$ est le nombre total de modalités, chaque individu est représenté par un vecteur ligne de M composantes à valeurs dans $\{0, 1\}$. Il n'y a qu'un 1 parmi les m_1 premières composantes, seulement un 1 entre la m_1+1-ième et la (m_1+m_2)-ième, etc

Pour simplifier, on peut numéroter toutes les modalités de 1 à M et noter Z_j, ($1 \le j \le M$) le vecteur colonne constitué des N réponses de la j-ème modalité. Le i-ième élément du vecteur Z_j vaut 1 ou 0, selon le choix de l'individu i (c'est 1 si et seulement si l'individu i a choisi la modalité j).

On peut ainsi définir une ($N \times M$) matrice D composée de 0 et 1 et dont les colonnes sont les vecteurs Z_j Elle est composée de K blocs où chaque ($N \times m_k$) bloc contient les N réponses de la question k. On a:

$$D = \left(Z_1, \cdots, Z_{m_1}, \cdots, Z_j, \cdots, Z_M\right)$$

La matrice D est appelé *tableau disjonctif complet* et on note

$$D = (d_{ij}),\ i = 1, ..., N,\ j = 1, ..., M.$$

Ce tableau D contient toute l'information concernant les individus. Il est le résultat brut de toute enquête.

	m_1			m_2		m_3		
Ind	1	2	3	1	2	1	2	3
1	0	1	0	0	1	1	0	0
2	1	0	0	1	0	0	1	0
…								
…								
i	0	0	1	0	1	0	0	1
N	0	0	1	1	0	0	0	1

Tableau 1: Exemple de Tableau Disjonctif Complet.

Si on veut savoir qui répond quoi, il est essentiel d'utiliser ce tableau. Mais si on veut seulement étudier les *relations entre les K variables (ou questions)*, on peut résumer les données de l'enquête dans une tableau croisé, *appelé table de Burt*, définie par

$$B = D'D$$

ou *D'* est la transposée de la matrice D. La matrice B de format ($M \times M$) est symétrique et est composée de $K \times K$ blocs, parmi lesquels le (k, l) bloc B_{kl} (pour $1 \le k \ne l \le K$) est le tableau de contingence qui croise la question k et la question l. Le bloc B_{kk} est une matrice diagonale, dont les termes sont les effectifs des individus qui ont choisi respectivement les modalités 1, ... , m_k, pour la question k.
La table de Burt peut être représentée comme ci-dessous. On peut la voir comme un tableau de contingence généralisé à plus de deux variables (ou questions).

A partir de maintenant, on note b_{jl} les entrées de la matrice B, ceci quelles que soient les questions associées aux modalités j ou l. L'entrée représente le nombre d'individus qui choisissent à la fois les modalités j et l. D'après les définitions des données, si j et l sont deux modalités différentes de la même question, $b_{jl} = 0$, et si $j = l$, l'entrée b_{jj} est le nombre d'individus qui choisissent la modalité j. Dans ce cas, on n'utilise qu'un seul indice et l'on écrit b_j à la place de b_{jj}. Ce nombre n'est rien d'autre que la somme des éléments du vecteur Z_j. Chaque ligne ou colonne de la matrice B caractérise une *modalité d'une question* (aussi appelée *variable*). Nous représentons ci-dessous la table de Burt associée au tableau disjonctif complet ($K = 3$, $m_1 = 3$, $m_2 = 2$ et $m_3 = 3$) reproduit dans le tableau 1 .

	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8
Z_1	b_1	0	0	b_{14}	b_{15}	b_{16}	b_{17}	b_{18}
Z_2	0	b_2	0	b_{24}	b_{25}	b_{26}	b_{27}	b_{28}
Z_3	0	0	b_3	b_{34}	b_{35}	b_{36}	b_{37}	b_{38}
Z_4	b_{41}	b_{42}	b_{43}	b_4	0	b_{46}	b_{47}	b_{48}
Z_5	b_{51}	b_{52}	b_{53}	0	b_5	b_{56}	b_{57}	b_{58}
Z_6	b_{61}	b_{62}	b_{63}	b_{64}	b_{65}	b_6	0	0
Z_7	b_{71}	b_{72}	b_{73}	b_{74}	b_{75}	0	b_7	0
Z_8	b_{81}	b_{82}	b_{83}	b_{84}	b_{85}	0	0	b_8

Tableau 2: Exemple de Table de Burt.

La somme totale des entrées de B est $b = \sum_{j,l} b_{jl} = K^2 N$

En effet,

sachant que pour chaque j on a $\sum_l b_{jl} = K b_j$ et que $\sum_j b_j = \sum_{k=1}^{K} \sum_{l=1}^{m_k} b_l = \sum_{k=1}^{K} N = KN$

on en déduit que

$$b = \sum_{j,l} b_{jl} = \sum_j \sum_l b_{jl} = \sum_j K b_j = K \sum_j b_j = K^2 N$$

Dans le prochain paragraphe, on présente l'Analyse des Correspondances Multiples qui est une méthode factorielle classique pour étudier les relations entre les modalités de différentes variables qualitatives.

4. Analyse Factorielle des Correspondances et Analyse des Correspondances Multiples

4.1 L'Analyse Factorielle des Correspondances

L'analyse des correspondances multiples (ACM) (Burt, 1950, Benzécri, 1973, Greenacre, 1984, Lebart et al., 1984) est une généralisation de l'Analyse Factorielle des Correspondances (AFC) qui traite un tableau de contingence. Considérons seulement deux variables qualitatives ayant respectivement I et J modalités. Le tableau de contingence pour ces deux variables est une matrice $I \times J$, 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 la table de Burt B peuvent être considérés comme des tableaux de contingence. En fait, D est le tableau de contingence qui croise la « méta-variable » INDIVIDU à N valeurs avec la « méta-variable » MODALITE de M valeurs. De la même façon, B est le tableau de contingence qui croise la « méta-variable » MODALITE de M valeurs avec elle-même.

Nous allons définir l'Analyse Factorielle des Correspondances, appliquée au tableau de contingence (Lebart et al., 1984).
On définit successivement

- 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

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

Comme 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 $f_{\bullet j}$.

Si on calcule l'inertie de ces deux ensembles de profils,

Inertie totale des profils lignes $= \sum_i f_{i\bullet} \chi^2(i,\bar{i})$

Inertie totale des profils colonnes $= \sum_j f_{\bullet j} \chi^2(j,\bar{j})$,

il est facile de voir que ces expressions sont égales. Cette inertie totale sera notée $\Im$

$$\Im = \sum_{i,j} \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{f_{i\bullet} f_{\bullet j}} = \sum_{i,j} \frac{f_{ij}^2}{f_{i\bullet} f_{\bullet j}} - 1$$

On peut souligner deux points importants :

1) Afin d'utiliser la distance euclidienne entre les lignes et entre les colonnes au lieu de la distance du chi-deux, et pour prendre en compte le poids de chaque ligne par $f_{i\bullet}$ et le poids de chaque colonne par $f_{\bullet j}$, il est très pratique de remplacer les valeurs initiales f_{ij} par les valeurs corrigées f_{ij}^C, en posant $f_{ij}^C = \frac{f_{ij}}{\sqrt{f_{i\bullet}}\sqrt{f_{\bullet j}}}$

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.

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.

D'après le paragraphe 1.3, la diagonalisation de la matrice $F^c{}'F^c$ peut être approximativement remplacée par l'algorithme SOM appliqué aux profils lignes, tandis que la diagonalisation de $F^cF^c{}'$ 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.

4.2. L'Analyse des Correspondances Multiples

Nous allons maintenant rappeler comment se définit l'algorithme classique des correspondances multiples.

1) **Si l'on s'intéresse uniquement aux modalités**, la table de Burt, considérée comme tableau de contingence, est prise en entrée. Comme on l'explique plus haut, on considère *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 $B^c{}'B^c$ ou $B^cB^c{}'$ sont identiques, et on obtient une représentation simultanée des M vecteurs lignes (i.e. des modalités) sur plusieurs plans factoriels, donnant une information sur les relations entre les K variables.

2) **Si l'on est intéressé par les individus**, il est nécessaire d'utiliser le tableau Disjonctif Complet, considéré lui aussi comme un tableau de contingence (voir plus haut).

Notons D^c la matrice corrigée, dont les entrées sont d_{ij}^c , données par

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

Dans ce cas, la matrice D^c n'est plus symétrique. La diagonalisation de $D^c{}'D^c$ donne une représentation des individus, celle de $D^cD^c{}'$ fournit une représentation des modalités. Les représentations modalités et individus peuvent être superposées. Dans ce cas, il est possible de calculer les coordonnées des modalités et des individus.

Concluons cette brève présentation de l'ACM par quelques remarques. C'est une méthode de projection linéaire, elle fournit plusieurs cartes bidimensionnelles, chacune d'entre elles représentant un faible pourcentage de l'inertie globale. Il est donc nécessaire de regarder plusieurs cartes à la fois, les modalités sont plus ou moins bien représentées, et il n'est pas toujours facile d'en déduire des conclusions pertinentes concernant la proximité entre les modalités. Des modalités liées sont projetées en des points voisins, mais du fait de la distorsion due à la projection, des points voisins peuvent ne pas correspondre à des modalités voisines. La principale propriété est que chaque modalité est placée approximativement au centre de gravité des modalités qui lui sont corrélées et des individus qui la partagent (si les individus sont disponibles). Mais cette approximation peut être très pauvre et les graphiques ne sont pas toujours faciles à interpréter, comme le montreront les exemples.

4.3. De l'Analyse des Correspondances Multiples à l'algorithme SOM

D'après la conclusion du paragraphe 1.3., 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), voir paragraphe 5.

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 (voir paragraphe 7).

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.

5. Analyse d'une table de Burt basée sur Kohonen : algorithme KACM

Dans ce paragraphe, on ne prend en compte que les modalités et on définit un algorithme fondé sur l'algorithme de Kohonen, qui est analogue à l'ACM sur une table de Burt.

Ces algorithmes ont été introduits par Cottrell, Letrémy, Roy (1993), Ibbou, Cottrell (1995), Cottrell, Rousset (1997), la thèse de Smaïl Ibbou (1998), Cottrell et al. (1999), Letrémy, Cottrell, (2003).

La matrice des données est la table de Burt corrigée B^c définie dans le paragraphe 4.2. On considère un réseau de Kohonen $n \times n$ (grille), muni de sa topologie habituelle.

Chaque unité u est représentée par un vecteur code C_u de $R^{\boldsymbol{M}}$; les vecteurs codes sont initialisés au hasard.
L'apprentissage à chaque étape consiste à
- présenter au hasard une entrée $r(j)$ i.e. une ligne de la *matrice corrigée* B^c,
- rechercher son unité gagnante u_0, i.e. celle qui minimise $\|r(j) - C_u\|^2$ pour toutes les unités u,
- mettre à jour les vecteurs codes (ou vecteurs poids) de l'unité gagnante et de ses voisines par

$$C_u^{new} = C_u^{old} + \varepsilon\sigma(u,u_0)\left(r(j) - C_u^{old}\right)$$

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. [1]

Après apprentissage, chaque profil ligne $r(j)$ est représenté par son unité gagnante. A cause de la propriété de conservation de la topologie de l'algorithme de Kohonen, la représentation des M entrées sur la grille $n \times n$ souligne la *proximité* entre modalités des K variables. Après convergence, on obtient une classification organisée de toutes les modalités, où les modalités liées appartiennent à la même classe ou à des classes voisines.

Nous appelons cette méthode KACM (Kohonen Analyse des Correspondances Multiples) ; elle fournit une très intéressante alternative à l'ACM classique.

6. Exemple I : La base de données des pays avec des variables qualitatives

On considère la base de données de POP_96 du paragraphe 2.3. Maintenant, nous considérons les 7 variables comme des variables qualitatives, en les discrétisant en classes, et nous ajoutons la variable IDH. La définition des 8 variables se trouve dans le tableau 3.

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

Tableau 3: Les variables qualitatives pour la base de données POP_96.

Il y a 8 variables qualitatives et 31 modalités.

Dans la Fig. 4, on représente le résultat de KACM appliqué à ces données. Nous pouvons constater que les « bonnes » modalités (terminant par 1) caractéristiques des pays développés sont toutes situées dans le coin inférieur gauche, suivies par les modalités « intermédiaires » (niveau 2) dans le coin supérieur gauche. Les « mauvaises » modalités sont situées dans le coin supérieur droit et les « très mauvaises » dans le coin inférieur droit. Trois classes vides séparent les meilleures modalités des moins bonnes.

[1] Le paramètre d'adaptation est défini par une fonction décroissante du temps t, qui dépend du nombre $n \times n$ des unités du réseau ($\varepsilon = \varepsilon_0 / (1 + c_0\ t\ /\ n \times n)$. Le rayon de voisinage est aussi une fonction décroissante de t, dépendant de n et de T_{max} (le nombre total d'itérations), $\rho(t) = \text{Integer}\left(\frac{n/2}{1+(2n-4)/T_{\max}}\right)$

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	

Fig. 4: La répartition des 31 modalités sur la carte de Kohonen avec KACM, 500 itérations.

Si nous examinons la projection (Fig.5) sur les deux premiers axes par une ACM, nous voyons à peu près la même disposition, avec un premier axe qui oppose les meilleures modalités (sur la gauche) aux plus mauvaises (sur la droite). La projection sur les axes 1 et 5 (Fig.6) permet une meilleure lisibilité mais avec une perte d'informations (seulement 30% d'inertie expliquée). Pour conserver 80% de l'inertie totale, il est nécessaire de considérer les dix premiers axes.

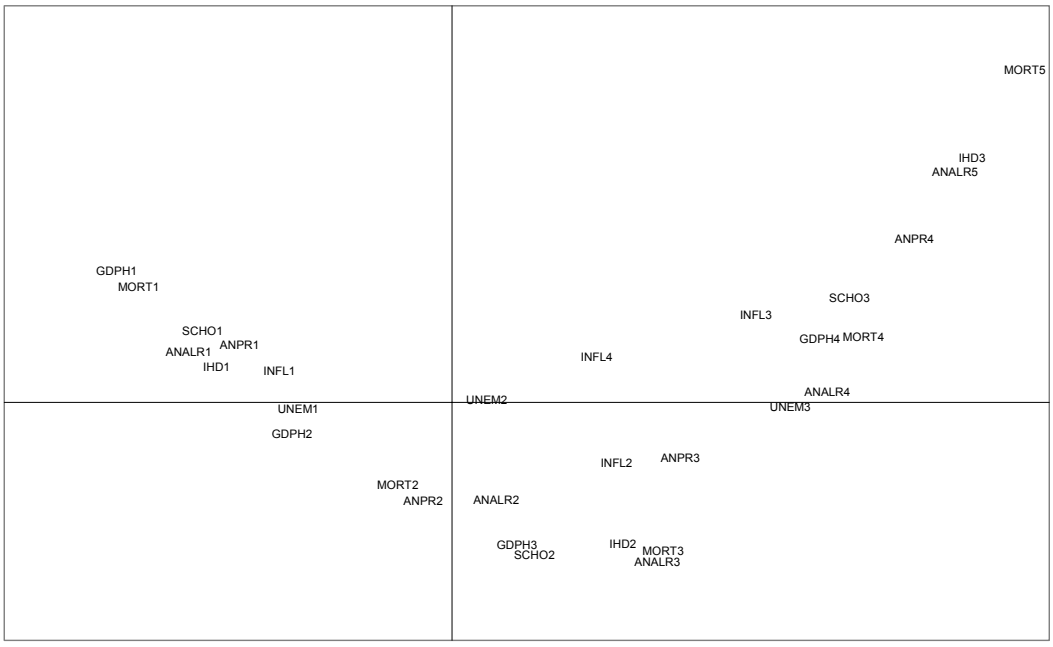

Fig 5: La représentation ACM, axes 1 (24%), et 2 (14%), 38% d'inertie expliquée.

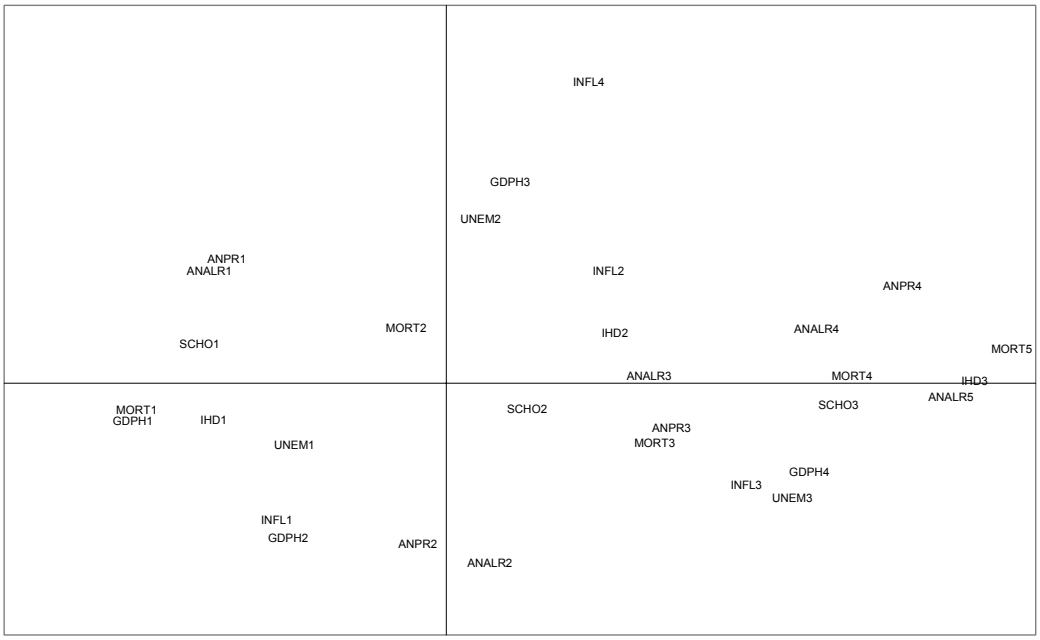

Fig 6: La représentation ACM, axes 1 (24%), et 5 (6%), 30% d'inertie expliquée.

Comme les modalités sont classées selon une échelle (irrégulière), on voit bien la progression du niveau 1 (le meilleur) aux moins bons (niveaux 3 et 4 ou 5 selon la modalité). Mais les clusters ne sont pas clairs. L'examen d'autres axes pourrait amener à des conclusions contradictoires : par exemple, les modalités INFL3 et INFL4 semblent proches dans la Fig.5 et éloignées dans la Fig.6.
On peut aussi réduire le nombre de classes comme dans le paragraphe 2.3 et rendre visible cette double classification. On attribue alors à chaque « macro-classe » une couleur ou un niveau de gris. Dans la Fig.7, on a regroupé les « micro-classes » en 6 « macro-classes ».

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	

Fig. 7: Macro-classes regroupant les modalités en 6 classes facilement interprétables.

7. Analyse des variables qualitatives en gardant les individus : KACM_ind

Jusqu'à présent, nous ne nous sommes intéressés qu'aux relations entre modalités. Mais il peut être plus intéressant et plus performant de regrouper ensemble les individus et les modalités qui les décrivent. Dans ce cas, il est évident qu'il faut utiliser le Tableau Disjonctif Complet afin de connaître les réponses individuellement.

Une manière facile de représenter simultanément les individus avec leurs modalités est de construire une carte de Kohonen avec les individus en utilisant l'algorithme SOM appliqué au Tableau Disjonctif Complet Corrigé et de projeter les modalités comme des données supplémentaires, avec une normalisation adaptée.

Par conséquent, on corrige le tableau D en D^c (voir paragraphe 4.2), l'algorithme SOM est appliqué aux lignes de ce tableau corrigé. Chaque modalité j est représentée par un M-vecteur, qui est le **vecteur moyen** de tous les individus partageant cette modalité.
Ses coordonnées sont:

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

avec les notations définies dans le paragraphe 3. 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.

KACM_ind nécessite plus d'itérations que KACM. Ceci est facilement compréhensible puisque nous classons d'abord les individus (qui sont généralement plus nombreux que les modalités) mais nous calculons aussi avec plus de précision les vecteurs codes qui sont utilisés comme prototypes pour assigner les modalités considérées comme données supplémentaires.

La visualisation n'est pas toujours utile, surtout lorsqu'il y a trop d'individus. Mais la carte peut montrer des associations entre modalités ou groupes de modalités et sous-ensembles d'individus.

On voit sur la Fig.8 la représentation simultanée des 96 pays et des 31 modalités. En appliquant une Classification Hiérarchique Ascendante, on les a regroupés en 7 macro-classes.

Fig. 8: Représentation simultanée des 96 pays et des 31 modalités, 20 000 itérations.

Nous observons que la proximité entre pays et modalités est parfaitement cohérente, les pays riches sont dans le coin inférieur gauche, avec presque toutes les « bonnes » modalités (niveau 1), les très pauvres dans le coin supérieur droit (niveaux 3,4 et 5), etc

Cette méthode permet une bonne représentation simultanée des modalités et des individus mais rompt la symétrie entre individus et modalités. Inversement, l'algorithme présenté dans le nouveau paragraphe conserve cette symétrie et est directement inspiré des méthodes classiques. Cet algorithme a été présenté par Cottrell et Letrémy (2003).

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

Pour garder ensemble les modalités et les individus de façon plus équilibrée, comme dans l'ACM classique, on définit un nouvel algorithme KDISJ qui est une extension de algorithme KORRESP (Cottrell et al., 1993) qui a été créé pour analysé des tableaux de contingence qui croisent deux variables qualitatives. L'algorithme KDISJ a d'abord été défini dans Cottrell, Letrémy, (2003). Une version étendue est présentée dans Cottrell, Letrémy (2004).

Le Tableau Disjonctif Complet est corrigé comme indiqué dans le paragraphe 4.2. On choisit ensuite un réseau de Kohonen, et on associe à chaque unité u 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). On note

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

pour mettre en évidence la structure du vecteur code C_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).

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

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

qui maximise le coefficient d_{ij}^c, c'est-à-dire la modalité la plus rare dans la population totale parmi les modalités qui lui correspondent. Ensuite, on crée un vecteur individu étendu $X = (i, j(i)) = (X_M, X_N)$, de dimension $(M + N)$ (voir Fig. 9). 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.

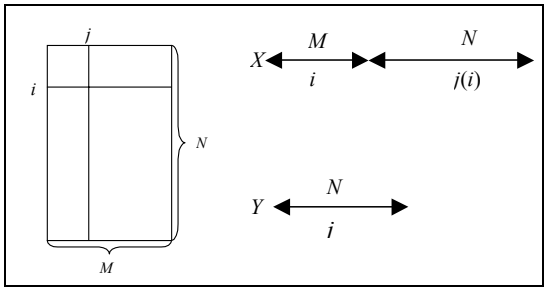

Fig. 9 : La matrice D^c, les vecteurs X et Y.

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), et σ est la fonction de voisinage telle que $\sigma(u, u_0)$ vaut 1 si u et u_0 sont voisines sur le réseau de Kohonen, et 0 sinon. Le paramètre d'adaptation et le rayon de voisinage ont été définis dans la note 1 du paragraphe 5.

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 (voir Fig. 9) le vecteur colonne de dimension N correspondant à la modalité j. Cette étape peut s'écrire :

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

où les M premières composantes ne sont pas modifiées. Voir dans la Fig. 10, le résultat de KDISJ appliqué à la base POP_96.

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	UNEM2 U Arab Emirates Malaysia	Malta Portugal	UNEM2 Greece Hungary Slovenia Uruguay	ANPR1 ANALR1 SCHO1 R. Czech
ANALR4 Comoros Ivory Coast	MORT4 Cameroon Nigeria	ANPR3 Bolivia	INFL2 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

Fig. 10: Carte de Kohonen avec représentation simultanée des modalités et des pays. Les 36 micro-classes regroupées en 7 macro-classes, 2 400 itérations.

On peut observer que les positions simultanées des pays et des modalités ont du sens, et que les macro-classes sont faciles à interpréter. Il est possible de contrôler le bon positionnement des modalités par rapport aux individus, en calculant les déviations à l'intérieur de chaque classe de Kohonen. La déviation pour une modalité l (partagée par b_l individus) et pour une classe k (avec n_k individus) peut être calculée comme la différence entre le nombre d'individus qui possède cette modalité et qui appartiennent à cette classe k et le nombre théorique $b_l\ n_k\ /\ N$ qui correspond à une situation d'indépendance entre modalité et

classification. Ces déviations sont toutes positives, ce qui signifie qu'il y a attraction entre les modalités et les individus qui appartiennent à la même classe.

On pratique ainsi un classement classique de Kohonen sur les individus, un autre sur les modalités, tout en les maintenant associés. Après convergence, les individus et les modalités sont classés dans les classes de Kohonen. Des individus ou modalités « proches » sont classés dans la même classe ou dans des classes voisines. On appelle KDISJ l'algorithme ainsi défini. Il suffit en général de faire 20(*M*+*N*) itérations pour obtenir la convergence.

On peut comparer la Fig. 10 avec les représentations que l'on obtient en utilisant l'ACM classique (voir Fig. 11 et 12) ; les caractéristiques de base sont les mêmes, mais dans les Fig. 11 et 12, les regroupements (clusters) ne sont pas clairement définis, il y a quelques contradictions d'une projection à l'autre (voir par exemple les localisations des modalités INFL3 et INFL4). La carte de Kohonen est beaucoup plus utile à la visualisation.

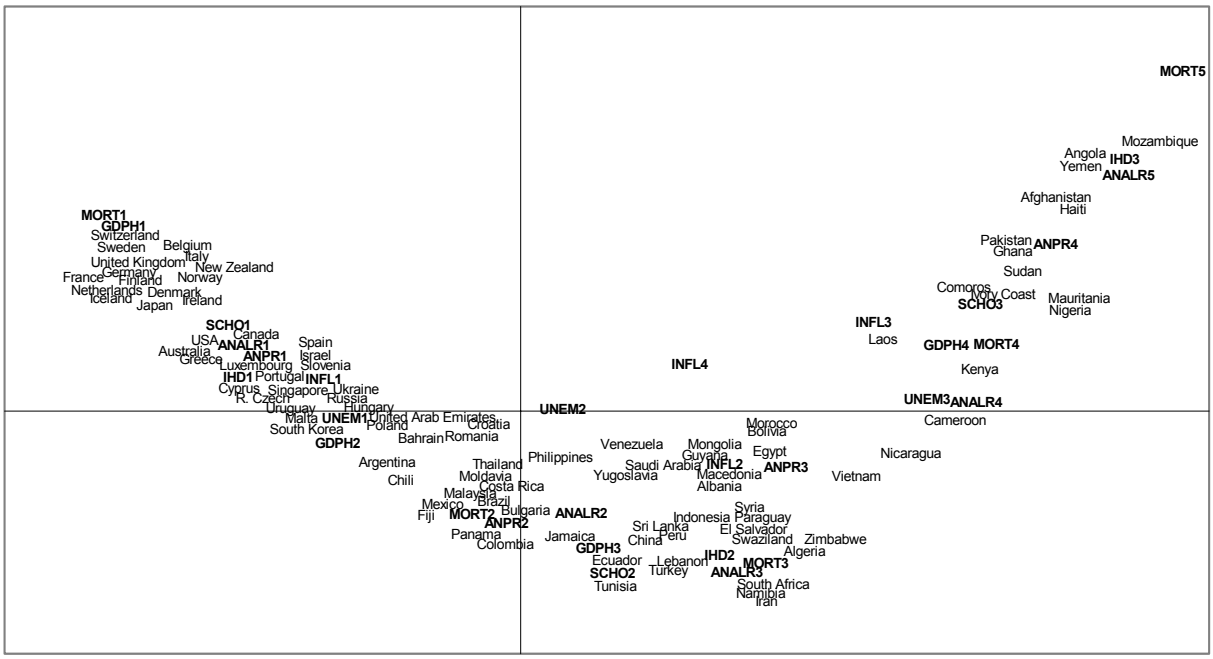

Fig. 11: Représentation ACM (modalités et individus), axes 1 (24%), et 2 (14%), 38% d'inertie expliquée.

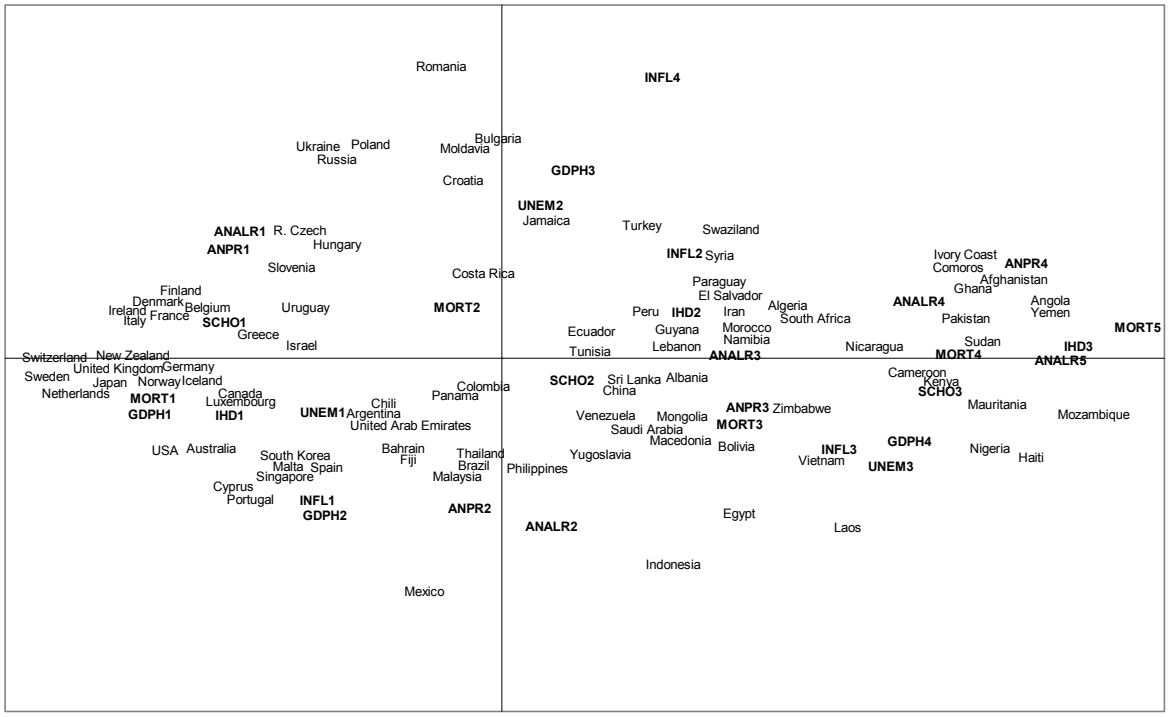

Fig. 12: Représentation ACM (modalités et individus), axes 1 (24%), et 5 (6%), 30% d'inertie expliquée.

9. Exemple II: 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 (voir tableau 4.) 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.

	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

Tableau 4: Tableau de contingence pour des couples mariés, (source INSE, 1990). Les lignes sont pour les hommes et les colonnes pour les femmes.

Dans ce qui suit, les couples sont notés (*i, j*) où $i = 1, ..., 6, j = 1, ..., 6$; *i* et *j* correspondent aux différentes catégories indiquées au-dessus.

On considère d'abord les résultats d'une ACM. On a besoin de 5 axes pour conserver 80% de l'inertie totale. La meilleure projection, sur les axes 1 et 2 de la Fig.13, est trop schématique. L'axe 1 oppose les fermiers à tous les autres couples mais déforme la représentation.

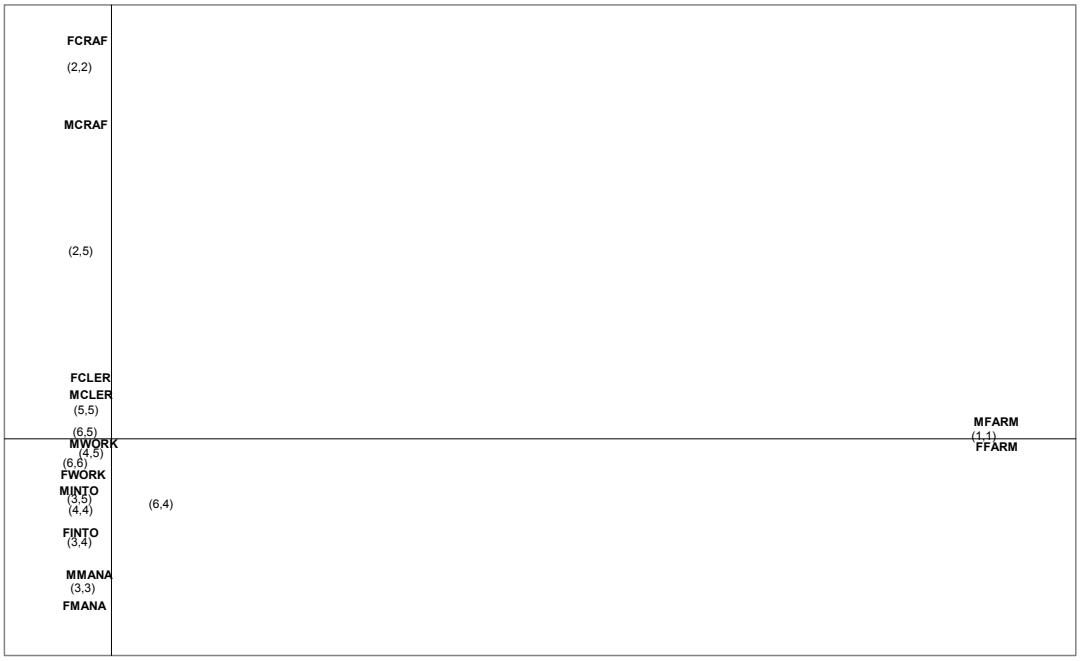

Fig.13 : La représentation ACM (modalités et individus), axes 1 (20%) et 2 (17%), 37% d'inertie expliquée

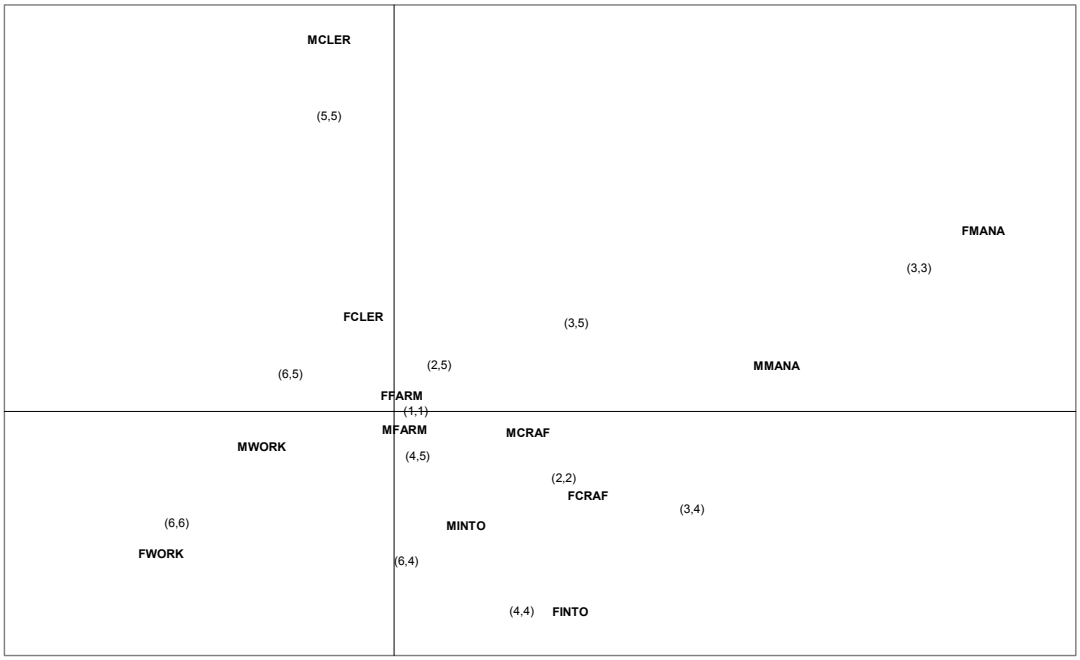

Fig 14 : La représentation ACM (modalités et individus), axes 3 (16%) et 5 (13%), 29% d'inertie expliquée

Sur les axes 3 et 5, la représentation est meilleure, même si le pourcentage d'inertie expliquée est plus faible. Dans les deux cas, chaque type de couple est mis exactement à mi-distance des modalités de chaque membre du couple.

Seuls 12 points sont visibles pour les individus, puisque les couples correspondant aux mêmes groupes professionnels pour le mari et la femme sont identiques (il n'y a pas d'autres variables).

On applique la méthode KACM à ces données très simples et on obtient la carte SOM de la Fig.15.

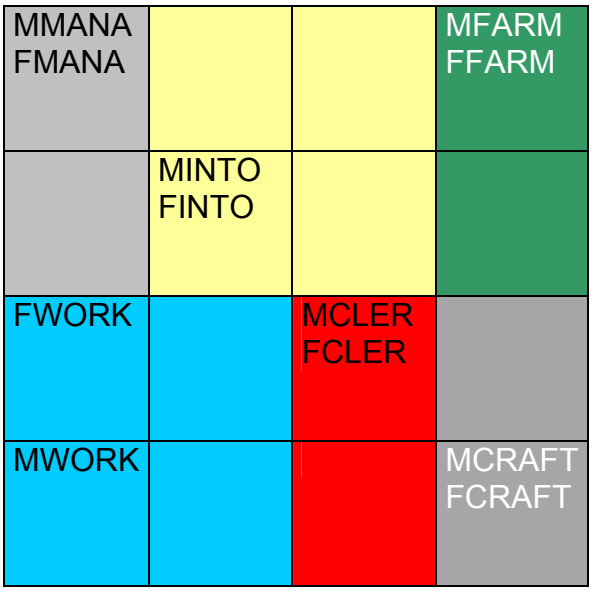

Fig.15 : Carte de Kohonen avec représentation des modalités. Les 16 micro-classes sont regroupées en 6 macro-classes qui regroupent uniquement des modalités identiques, 200 itérations.

Dans ce qui suit (Fig.16), nous utilisons l'algorithme KACM_ind. Nous indiquons le nombre de chaque type de couple présent dans chaque classe.

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)

Fig.16 : KACM_ind : carte de Kohonen avec représentation simultanée des modalités et des individus. Le nombre de couples de chaque type est indiqué. 16 (1,1) signifie qu'il y a 16 couples où le mari et la femme sont tous les deux fermiers, 10000 itérations.

Dans la Fig.17, nous représentons les résultats obtenus avec l'algorithme KDISJ. Les résultats sont tout à fait similaires. Dans les deux cartes de Kohonen, chaque type de couple est situé dans la même classe que les catégories professionnelles des deux membres du couple, ou entre les modalités correspondantes.

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)

Fig.17 : KDISJ : Carte de Kohonen avec représentation simultanée des modalités et des individus. Le nombre de couples de chaque type est indiqué. 16 (1,1) signifie qu'il y a 16 couples où le mari et la femme sont tous les deux fermiers, 5 000 itérations.

Dans cet exemple jouet, nous pouvons par conséquent conclure que les résultats sont tout à fait satisfaisants, puisqu'ils donnent les mêmes informations que les projections linéaires, avec l'avantage qu'une seule carte suffit pour résumer la structure des données.

10. Exemple III : les emplois intérimaires

Dans ce paragraphe, nous présentons un autre exemple tiré d'une vaste étude du temps de travail en 1998-1999 faite par l'INSEE à partir d'une enquête. Le rapport (Letrémy, Macaire et al., 2002) cherche à étudier les « formes particulières d'emplois (FPE)**»,** à savoir i) les CDD, ii) les emplois à temps partiel, à durée déterminée ou indéterminée, iii) 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 …
Une étude initiale intitulée « Temps de travail dans les FPE : le cas du travail à temps partiel » (Letrémy, Cottrell, 2003) couvrait 14 des questions, correspondait à 39 modalités de réponse et concernait 827 travailleurs à temps partiel.

Dans ce paragraphe, nous nous intéressons aux travailleurs intérimaires ; ils sont 115. Nous présentons une application des algorithmes KACM et KDISJ utilisés aussi bien pour classer les modalités que les individus.

Le tableau 5 recense les variables et les modalités de réponse. Il y a 25 modalités.

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

Tableau 5 : variables utilisées dans l'enquête individuelle

Dans la Fig. 18, les modalités sont représentées sur une carte de Kohonen, après avoir été classées selon un algorithme KACM.

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

Fig 18 : Représentation des modalités regroupées en 6 clusters, 500 itérations

Les clusters sont clairement identifiables : les meilleures conditions de travail se situent dans le coin inférieur droit (niveau n°1), les plus jeunes (AGE1) sont dans le coin inférieur gauche auquel correspondent de mauvaises conditions de travail (travail le samedi, de nuit…)

On retrouve les mêmes associations sur la représentation ACM, voir Fig.19 (axes 1 et 2).

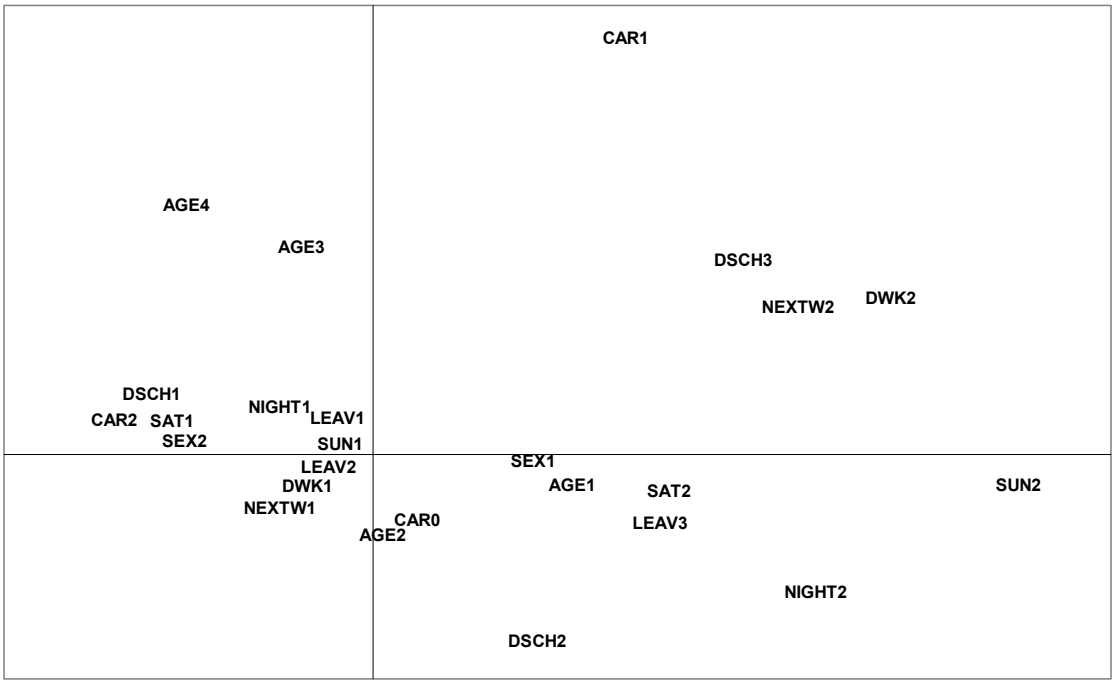

Fig 19 : La représentation ACM, axes 1 (19%) et 2 (11%), 30% d'inertie expliquée

Maintenant, nous pouvons classer en même temps les modalités et les individus en utilisant l'algorithme KDISJ. Voir Fig.20.

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

Fig.20 : KDISJ : Classement simultané de 25 modalités et de 115 individus. Seul le nombre d'individus est indiqué dans chaque classe, 3000 itérations.

Les groupes sont faciles à interpréter, les bonnes conditions de travail sont regroupées, ainsi que les mauvaises.

Comme d'habitude, la projection simultanée des individus et des modalités sur les deux premiers axes d'une ACM n'est pas claire et la visualisation est pauvre. Sur la Fig. 21, seulement 30% de l'inertie totale est prise en compte et il faudrait 9 axes pour obtenir 80% de l'inertie totale.

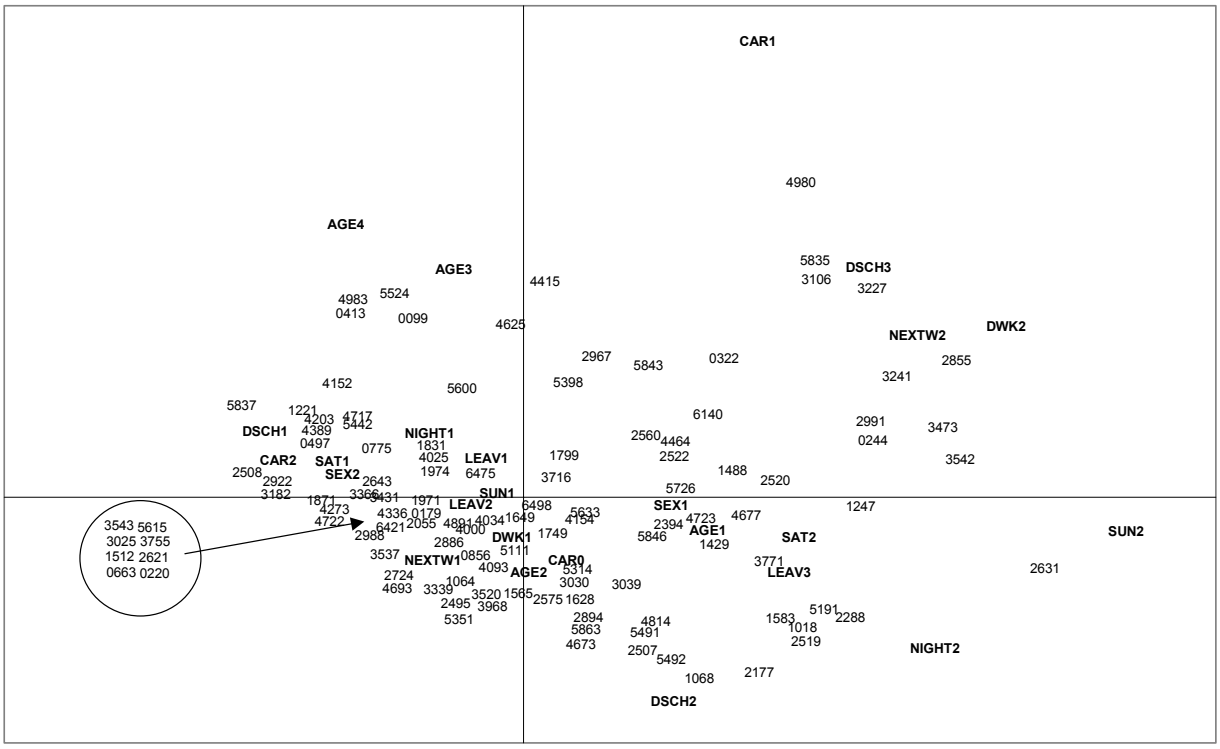

Fig.21 : Représentation ACM simultanée des 25 modalités et des 115 individus, axes 1 (19%) et 2 (11%), 30% de l'inertie expliquée.

11. Conclusion

Nous proposons plusieurs méthodes pour analyser des données multidimensionnelles, en particulier lorsque les observations sont décrites par des variables qualitatives, en complément des méthodes linéaires et factorielles classiques. Ces méthodes sont des adaptations de l'algorithme original de Kohonen.

Dans le tableau 6, 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	

Tableau 6 : Comparaison entre les méthodes SOM et les méthodes factorielles

Mais en fait, pour les applications, il est nécessaire de combiner différentes techniques. Par exemple, lorsque les variables sont quantitatives, il est souvent intéressant de réduire d'abord la dimension en utilisant une ACP et en ne conservant qu'un nombre réduit de coordonnées.

D'un autre côté, s'il y a à la fois des variables qualitatives et quantitatives, il est utile de construire une classification des observations en ne retenant que les variables quantitatives, en utilisant une classification de KOHONEN suivie d'un Algorithme Hiérarchique Ascendant afin de définir une nouvelle variable qualitative. Elle vient s'ajouter aux autres variables qualitatives et on peut appliquer une Analyse des Correspondances Multiples ou un KACM à toutes les variables qualitatives (les variables de départ et la nouvelle variable). Cette technique permet une interprétation facile des classes et montre bien la proximité entre les modalités.

Si l'on s'intéresse aux seuls individus, on peut transformer les variables qualitatives en variables réelles grâce à une Analyse des Correspondances Multiples, auquel cas on garde tous les axes et on peut alors décrire chaque observation par ses coordonnées. La base de données devient alors numérique et on peut l'analyser grâce à un algorithme classique de classification ou par un algorithme de Kohonen.

Dans cet article, nous ne donnons pas d'exemple de carte de Kohonen à une dimension. Mais lorsqu'il est utile d'établir un score des données, la construction d'une ficelle de Kohonen (de dimension 1) à partir des données ou des vecteurs codes établis à partir des données, donne directement un score en « ordonnant » les données.

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 le data mining.

Références bibliographiques

Anderberg, M.R. 1973. *Cluster Analysis for Applications*, New-York, Academic Press.

Benzécri, J.P. 1973. *L'analyse des données, T2, l'analyse des correspondances,* Dunod, Paris.

Blayo, F. & Demartines, P. 1991. Data analysis : How to compare Kohonen neural networks to other techniques ? In *Proceedings of IWANN'91*, Ed. A.Prieto, Lecture Notes in Computer Science, Springer-Verlag, 469-476.

Blayo, F. & Demartines, P. 1992. Algorithme de Kohonen: application à l'analyse de données économiques. *Bulletin des Schweizerischen Elektrotechnischen Vereins & des Verbandes Schweizerischer Elektrizitatswerke*, 83, 5, 23-26.

Blayo F., Verleysen M., 1996. : *Introduction aux réseaux de neurones artificiels,* Collection Que-sais-je ?, PUF.

Burt, C. 1950. The factorial analysis of qualitative data, *British Journal of Psychology*, 3, 166-185.

Cottrell, M., Letrémy, P., Roy E. 1993. Analyzing a contingency table with Kohonen maps : a Factorial Correspondence Analysis, *Proc. IWANN'93*, J. Cabestany, J. Mary, A. Prieto (Eds.), Lecture Notes in Computer Science, Springer-Verlag, 305-311.

Cottrell, M. & Rousset, P. 1997. The Kohonen algorithm : a powerful tool for analysing and representing multidimensional quantitative et qualitative data, *Proc. IWANN'97*, Lanzarote. *Juin 1997*, J.Mira, R.Moreno-Diaz, J.Cabestany, Eds., Lecture Notes in Computer Science, n° 1240, Springer, 861-871.

Cottrell, M., Gaubert, P., Letrémy, P., Rousset P. 1999. Analyzing and representing multidimensional quantitative and qualitative data: Demographic study of the Rhöne valley. The domestic consumption of the Canadian families, *WSOM'99,* In: Oja E., Kaski S. (Eds), *Kohonen Maps*, Elsevier, Amsterdam, 1-14.

Cottrell, M. & Letrémy P. 2003. Analyzing surveys using the Kohonen algorithm, *Proc. ESANN 2003*, Bruges, 2003, M.Verleysen Ed., Editions D Facto, Bruxelles, 85-92.

Cottrell, M., Ibbou, S., Letrémy, P., Rousset, P. 2003 « Cartes auto-organisées pour l'analyse exploratoire de données et la visualisation » Journal de la Société Française de Statistique, Vol. 144, n°4, p67-106.

Cottrell, M. & Letrémy P. 2004. How to use the Kohonen algorithm to simultaneously analyze individuals and modalities in a survey, *à paraître dans Neurocomputing*.

Cottrell, M., Ibbou, S. Letrémy P. 2004. SOM-based algorithms for qualitative variables, *à paraître dans Neural Network*.

De Bodt, E., Cottrell, M., Letrémy, P., Verleysen, M. 2003. On the use of self-organizing maps to accelerate vector quantization, *Neurocomputing,* 56, 187-203.

Greenacre, M.J. 1984. *Theory and Applications of Correspondence Analysis*, London, Academic Press.

Ibbou, S. & Cottrell, M. 1995. Multiple correspondence analysis of a crosstabulation matrix using the Kohonen algorithm, *Proc. ESANN'95*, M.Verleysen Ed., Editions D Facto, Bruxelles, 27-32.

Ibbou, S. 1998. Classification, analyse des correspondances et méthodes neuronales, PHD Thesis, Université Paris 1.

Kaski, S. 1997. Data Exploration Using Self-Organizing Maps, *Acta Polytechnica Scandinavia*, *Mathematics, Computing and Management in Engineering Series* No. 82, (D.Sc. Thesis, Helsinki, University of Technology)

Kohonen, T. 1984, 1993. *Self-organization and Associative Memory*, 3°ed., Springer.

Kohonen, T. 1995, 1997. *Self-Organizing Maps*, Springer Series in Information Sciences,² Vol 30, Springer.

Lance, G.N. & Williams, W. T. 1967. A general Theory of Classification Sorting Strategies, *Computer Journal,* Vol. 9, 373-380.

Lebart, L., Morineau, A., Warwick, K.M. 1984. *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices,* Wiley.

Lebart, L., Morineau, A., Piron, M. 1995. *Statistique exploratoire multidimensionnelle*, Paris, Dunod.

Letrémy, P., Cottrell, M., Macaire, S., Meilland, C., Michon, F. 2002. Le temps de travail des formes particulières d'emploi, Rapport final, IRES, Noisy-le-Grand, February 2001, *Economie et Statistique*, Octobre 2002.

Letrémy, P. & Cottrell, M. 2003. Working times in atypical forms of employment : the special case of part-time work, *Conf. ACSEG 2001*, Rennes, 2001, *in* Lesage C., Cottrell M. (eds), *Connectionnist Approaches in Economics and Management Sciences*, Kluwer (Book Series: Advances in Computational Management Science, Volume 6), Chapter 5, p. 111-129.

Ripley, B.D. 1996. *Pattern Recognition and Neural Networks*, Cambridge University Press.

Rumelhart, D.E., McClelland, J.L. (eds). 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1. Foundations.* Cambridge, MA: The MIT Press.

Saporta, G. 1990. *Probabilités, Analyse de données et Statistique*, Paris, Technip.