

REVUE DE STATISTIQUE APPLIQUÉE

PATRICK ROUSSET

CHRISTIAN GUINOT

**Visualisation des distances entre les classes de la
carte de Kohonen pour le développement d'un outil
d'analyse et de représentation des données**

Revue de statistique appliquée, tome 50, n° 1 (2002), p. 35-47

http://www.numdam.org/item?id=RSA_2002__50_1_35_0

© Société française de statistique, 2002, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

VISUALISATION DES DISTANCES ENTRE LES CLASSES DE LA CARTE DE KOHONEN POUR LE DÉVELOPPEMENT D'UN OUTIL D'ANALYSE ET DE REPRÉSENTATION DES DONNÉES

Patrick ROUSSET^{(1),(2)}, Christian GUINOT⁽³⁾

⁽¹⁾ CEREQ, Service Informatique, 10 place de la Joliette 13474 Marseille cedex,

⁽²⁾ SAMOS, Université de Paris 1, 90 rue de Tolbiac 75013 Paris

⁽³⁾ CE.R.I.E.S., 20 rue Victor Noir, 92521 Neuilly sur Seine cedex

RÉSUMÉ

L'algorithme de Kohonen a la particularité de fournir une méthode de classification couplée avec une représentation appelée *carte de Kohonen* qui traduit la topologie entre les classes. De nombreux outils visuels d'aide à l'interprétation de cette classification, par exemple à partir d'une variable qualitative exogène ou endogène, prennent pour support cette carte afin de tenir compte du voisinage entre les classes. Dans cet article, nous proposons un outil qui visualise la distance entre toutes les classes et permet de rendre compte de la structure intrinsèque du nuage de points. L'algorithme de Kohonen devient ainsi le support d'une nouvelle technique d'analyse de données multidimensionnelles intégrant des représentations graphiques. En particulier, sa carte associée peut être un support de visualisation ou de comparaison de plusieurs classifications d'une même base de données (par exemple résultant des méthodes de segmentation et de classification hiérarchique). Cette méthode non linéaire peut ainsi suppléer le couple classique *analyse factorielle-classification* quand ce dernier n'est pas satisfaisant. La carte de Kohonen peut aussi être perçue comme la représentation d'une surface passant par les centres des classes associées. L'algorithme de Kohonen fournit alors une technique d'ajustement et de représentation du nuage de points.

Mots-clés : Algorithmes d'auto-organisation, Analyse de données, Cartes de Kohonen, Classification, Représentation de données, Réseau de neurones.

ABSTRACT

The Kohonen's self-organising map (SOM) is a classification algorithm whose associated map is a representation of the result reflecting the topology between classes. Many visual tools designed to help the interpretation are based on this map in order to take into account the neighbourhood structure between classes, for example one represents exogenous or endogenous qualitative variable effects on the classification. A new tool visualises distances between all the classes and so permits to show the data intrinsic structure. The SOM becomes then a multidimensional data analysis method which products some representations. In particular, the associated map permits to visualise or compare any classification on the same data set (for example resulting of clustering and segmentation methods). So, this non linear method can be used instead of the classical combination *factorial analysis – classification* when this

last one is not satisfying. The Kohonen map can be also considered as the representation of a surface which joins the Kohonen classes centroids. The SOM is then a new technique of data set adjustment associated with a representation.

Keywords : *Classification, Data set visualisation, Kohonen maps, Multivariate data analysis, Neural networks, Self organising maps.*

1. Introduction

La représentation de l'information fournie par les méthodes de classification est souvent peu satisfaisante. Un outil visuel qui permettrait à la fois de désigner les classes voisines entre elles et de comprendre la disposition de ces classes dans l'espace d'entrée pourrait améliorer fortement l'interprétation des classifications. Il permettrait par exemple de travailler confortablement sur un nombre de classes important ou de superposer l'information issue de deux classifications complémentaires, par exemple provenant de deux approches différentes telles que regroupement ou discrimination.

Actuellement, coupler une méthode de classification avec une analyse factorielle est très souvent utilisé (Wong 1982, Lebart *et alt.* 1995), mais le phénomène d'écrasement dû aux projections sur les plans factoriels rend ces représentations insatisfaisantes et leur interprétation risquée. La classification par l'algorithme de Kohonen (Kohonen 1993, Kohonen 1995, Cottrell *et alt.* 1998) offre l'originalité de proposer une visualisation de la structure de voisinage entre les classes. Il est de plus intéressant de noter que contrairement à la représentation obtenue avec le couple *analyse factorielle-classification*, celle produite par l'algorithme de Kohonen est conçue pour optimiser l'exploitation visuelle des propriétés de l'algorithme, c'est-à-dire sa notion de voisinage entre les classes. Plusieurs utilisations des cartes de Kohonen ont été proposées pour aider à l'interprétation de la classification (Cottrell *et alt.* 1997, Rousset 1999), malheureusement elles ne permettent pas de visualiser la structure de cette carte dans l'espace d'entrée. Ceci est très regrettable quand on veut utiliser les cartes de Kohonen comme outil d'analyse de données (Blayo *et alt.* 1991, Cottrell *et alt.* 1995, Rousset 1999) et particulièrement pour représenter le nuage de points. Un outil capable de visualiser les distances entre toutes les classes, et pas seulement les classes voisines, peut résoudre ce problème et permettre d'exploiter les cartes de Kohonen à la fois pour résumer les données et représenter le nuage de points. Les cartes de Kohonen, complétées par cet outil de visualisation des distances entre les classes, deviennent donc une technique d'analyse et de représentation de données capable de se substituer aux plans de l'analyse factorielle lorsque ceux-ci ne donnent pas satisfaction. On peut, en particulier, l'utiliser pour représenter le résultat de classifications issues d'autres méthodes et situer les classes des différentes classifications les unes par rapport aux autres dans l'espace d'entrée.

En effet, si les cartes de Kohonen n'ont été conçues initialement que pour optimiser la représentation de la classification de Kohonen, on peut espérer qu'elles soient également efficaces pour représenter des classifications issues d'autres méthodes qui utilisent la même distance. A titre d'exemple, on propose de superposer l'information de deux classifications de la même base de données, toutes deux faites avec la distance euclidienne, à partir de deux approches différentes. La première approche est une classification hiérarchique de type Ward qui fournit un rassemblement des indi-

vidus semblables pour chaque niveau de regroupement, la deuxième est une approche de type discriminante qui aboutit à un arbre de segmentation. Ces deux approches ont été utilisées sur les données d'une étude menée au C.E.R.I.E.S. dont l'objectif était de proposer une typologie de la peau humaine saine reposant sur un petit nombre de caractéristiques cutanées pertinentes (Guinot *et alt.* 1997, Chavent *et alt.* 1999). Les données ont été recueillies entre avril et mai 1996, sur 212 femmes volontaires d'Ile-de-France présentant une peau apparemment saine et d'âge compris entre 20 et 50 ans. Ces données résultent d'un examen médical appréciant 17 caractéristiques de la peau de la joue, évaluées sur des échelles qualitatives. Ces caractéristiques cutanées peuvent être visuelles comme «l'aspect gras de la peau» ou encore «la couleur jaune de la peau». Elles peuvent également être tactiles comme «la sensation rêche au toucher» ou encore «l'incapacité à rosir même après un léger pincement». Ces variables sont toutes binaires ou ordinales.

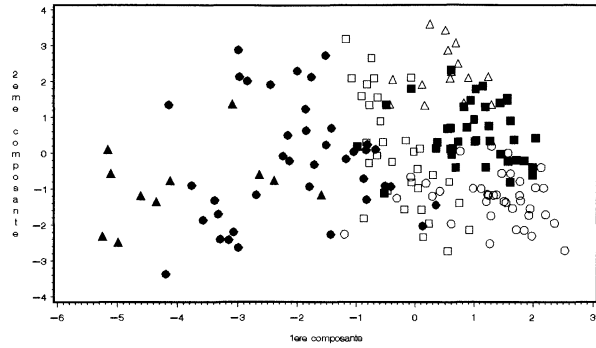
Après avoir présenté les cartes de Kohonen comme outil d'analyse puis de représentation des données et les avoir utilisées pour comparer plusieurs classifications, nous préciserons de nouvelles directions en vue de les exploiter pour fournir de nouveaux types d'ajustements des nuages de points, en particulier par des surfaces non linéaires, qui pourraient être visualisés par les représentations graphiques que nous présentons ici.

2. Représentation grâce au couple analyse factorielle-classification

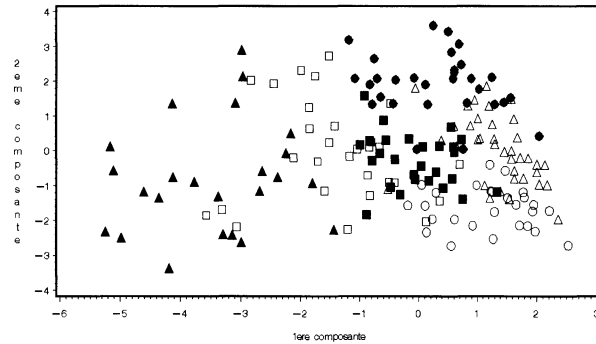
Une méthode classique de représentation des résultats d'une classification consiste à coupler celle-ci avec une analyse factorielle. La projection de tous les individus sur les plans principaux s'effectue en pointant ceux-ci avec un symbole qui indique leur classe d'affectation. Cette méthode, très utilisée, présente néanmoins les inconvénients dus aux projections. En effet, en utilisant cette technique, on explique la classification à partir d'un espace vectoriel plus petit que celui qui a été utilisé pour construire la classification. Il y a donc un problème de représentativité de la projection qui se traduit graphiquement par un phénomène d'écrasement bien connu. Si l'espace de projection ne concentre pas toute l'information qui a déterminé les classes, l'interprétation des proximités entre les classes risque d'être erronée. En fait, cette méthode est plus fiable quand on veut expliquer les axes principaux par les classes, ce qui n'est pas l'objectif ici.

Pour illustrer ces phénomènes, on a projeté les individus répartis dans les classes issues de la classification hiérarchique avec la distance de Ward dans notre exemple de typologie de la peau humaine saine (figure 1a). Le premier axe principal distingue les classes 2 ● et 6 ▲ des autres classes vers les valeurs négatives et le second axe les classes 1 ○ et 5 △. Si cette technique est une aide intéressante pour visualiser le positionnement des classes sur le plan, on peut néanmoins constater que cette projection superpose la classe 3 □ avec les classes 1 ○ et 2 ●. De plus, les classes 1 ○ et 6 ▲ qui apparaissent loin des autres sur ce plan ne le sont pas nécessairement dans l'espace complet. Le même type de difficulté est rencontré pour la classification issue de la segmentation (figure 1b).

Dans le cas particulier où l'on veut comparer deux classifications C_1 et C_2 , on peut traduire leur croisement par un tableau de contingence. Une identité des



(a)



(b)

FIGURE 1

- Projection sur le premier plan principal résultant d'une analyse en composantes principales : (a) individus représentés par le code de leurs classes issues de la classification hiérarchique avec la distance de Ward (classe 1 ○; classe 2 ●; classe 3 □; classe 4 ■, classe 5 △; classe 6 ▲)*
(b) individus représentés par le code de leurs groupes issus de la segmentation (classe 1 ○; classe 2 ●; classe 3 □; classe 4 ■, classe 5 △; classe 6 ▲)

deux classifications crée alors un tableau où seule la diagonale est non nulle (à une permutation des colonnes près). Dans le cas contraire, la projection sur le plan factoriel peut constituer une aide pour distinguer deux phénomènes distincts possibles. Le premier est le cas où une classe c de C_1 se répartit en deux classes issues de C_2 voisines dans l'espace et le second est celui où c se divise en deux classes distantes dans l'espace. Dans le deuxième cas, on conclut à une divergence entre les deux approches méthodologiques. Mais le problème des projections peut ici aussi fausser notre perception de la situation.

Dans notre exemple sur la typologie de la peau humaine saine, le tableau 1, construit à partir du croisement entre les classes de la classification hiérarchique (appelées classes) et celles de la segmentation (appelées groupes) constitue un support pour évaluer la cohérence entre les deux approches. Il fait apparaître une similarité

TABLEAU 1

Répartition des sujets dans les classifications résultant des deux approches.
Tableau de contingence croisant les classes issues des deux classifications.
Les effectifs sont exprimés en pourcentage du nombre total d'individus.

Classification hiérarchique (distance de Ward)	Classification issue de la segmentation					
	Groupe 1 ○	Groupe 5 △	Groupe 2 ●	Groupe 3 □	Groupe 4 ■	Groupe 6 ▲
Classe 1 ○	14,2 %	0,5 %	2,4 %	8,5 %	0,5 %	
Classe 2 ●		12,7 %				6,6 %
Classe 3 □	3,3 %		8 %		4,3 %	
Classe 4 ■	0,9 %	0,9 %	2,8 %	15,1 %	2,4 %	
Classe 5 △					11,3 %	
Classe 6 ▲		0,5 %				5,2 %

d'ensemble visible sur la diagonale. Néanmoins, les classes 3 □, 1 ○ et 2 ● se répartissent respectivement entre les trois ensembles de groupes «1 ○, 2 ● et 4 ■», «1 ○, 2 ● et 3 □» et «5 △ et 6 ▲». Ceci pose le problème de savoir si la classe 3 □ est constituée de trois sous-groupes ou si les groupes 1 ○, 2 ● et 4 ■ sont suffisamment proches pour avoir contribué à la constitution d'une classe 3 □ homogène. On constate que la projection sur le premier plan principal (figures 1a et 1b) ne permet pas de répondre à cette question.

3. Classification de Kohonen et la carte associée

On s'intéresse dans ce paragraphe à l'algorithme de Kohonen comme outil de classification adapté à toutes les distances. Dans la suite pour éviter les ambiguïtés on attribuera le préfixe *k* à ce qui concerne cette technique (par exemple *k*-classes). Les caractéristiques de cet algorithme d'auto-organisation non supervisé le rapprochent de la famille des algorithmes de nuées dynamiques, en particulier son nombre de classes est fixé au départ. Il est même identique à l'algorithme de Lloyd (Lloyd 1982) dans la phase finale de son apprentissage (appelé «apprentissage à 0 voisin»).

Le résultat de l'apprentissage est la détermination de *U* vecteurs appartenant à l'enveloppe convexe des données et appelés vecteurs codes. Ces vecteurs constituent un résumé de l'information contenue dans la base des données. A chaque individu est affecté le vecteur code le plus proche au sens de la distance sur l'espace des données. Deux individus associés au même vecteur code sont regroupés dans une même classe. On définit ainsi *U* *k*-classes dont les représentants dans l'espace des données sont les vecteurs codes associés. On peut noter que lorsque l'on fait tendre la durée de la dernière phase de l'apprentissage (à 0 voisin) vers l'infini, le vecteur code tend vers le centre de la *k*-classe, c'est-à-dire son barycentre. Cet algorithme est associé à un ensemble de *U* unités disposées en réseau pour constituer la carte appelée *carte de*

Kohonen. Chaque unité est affectée à un vecteur code et représente sur la carte la k -classe associée. Ces unités ont une disposition sur la carte qui traduit la proximité entre les vecteurs codes associés. Deux unités voisines sur la carte correspondent à deux vecteurs codes voisins au sens de la distance définie sur l'espace des données. La structure du réseau est libre, nous présenterons le cas particulier d'une grille carrée où les unités voisines d'une unité u_0 sont : au rayon 0 l'unité u_0 elle-même, au rayon 1 le carré de 9 unités centré sur u_0 , au rayon 2 le carré de 25 unités centré sur u_0 . Ce type de représentation permet de visualiser certaines caractéristiques de la k -classification. Elle offre aussi l'avantage de contrôler visuellement si les propriétés d'une k -classe s'étendent à ses voisines. Par exemple, on peut montrer graphiquement l'effet d'une variable qualitative Q ayant participé ou non à la construction de la k -classification en insérant dans chaque unité de la carte un camembert dont chaque tranche représente la fréquence d'une modalité de la variable Q dans la k -classe associée à cette unité.

La figure 2a représente la caractérisation de chaque k -classe à partir de la caractéristique cutanée «couleur jaune de la peau» dans le cadre de notre exemple sur la typologie de la peau humaine saine. Les unités étant numérotées par ordre croissant de gauche à droite et de haut en bas, on distingue aisément que la présence de cette caractéristique (matérialisée par les parts sombres des camemberts) est une particularité des individus affectés aux deux régions de k -classes voisines – unités 1, 2 et 8 – et – unités 28, 34, 35, 41, 42, 48 et 49 –, la région – 7, 14, 21 – étant mixte.

L'inconvénient majeur de cette représentation est que les unités sont représentées à distances égales ce qui n'est pas le cas de leurs vecteurs codes associés dans l'espace des données. On perd ainsi toute idée de la structure du nuage de points. Dans l'exemple de la typologie de la peau humaine saine, cela se traduit par une incapacité à voir sur la carte si les deux régions caractérisées par la présence «d'une couleur jaune de la peau» en haut à gauche et en bas à droite sont proches ou éloignées dans l'espace des données. Ce problème de la visualisation de la structure de l'espace des données va être abordé dans la section suivante.

4. Visualisation de la structure du nuage de points par une représentation de la distance entre les centres de classes

Pour visualiser la structure du nuage de points, ou plus modestement l'importance de la proximité entre les k -classes, une première idée (Cottrell-Rousset 1997, Rousset 1999) consiste à regrouper les vecteurs codes par une nouvelle classification, par exemple de type hiérarchique avec la distance de Ward. On appelle les nouvelles classes macro-classes. Un niveau de gris est affecté derrière chaque camembert à chaque macro-classe pour les distinguer sur la carte. En pratique, quand on applique cette technique, on constate le plus souvent que les macro-classes confirment la topologie de voisinage en regroupant des k -classes voisines sur la carte. Sur l'exemple de la typologie de la peau humaine saine (figure 2b), la macro-classe située dans l'angle en haut à gauche (indiquée par le deuxième niveau de gris) regroupe un nombre de k -classes (4) beaucoup plus petit que celle localisée dans l'angle en bas à gauche (indiquée par le troisième niveau de gris) (8). De plus, comme sur la figure 2a présentée dans la section 3, il n'est toujours pas possible de conclure que les angles en

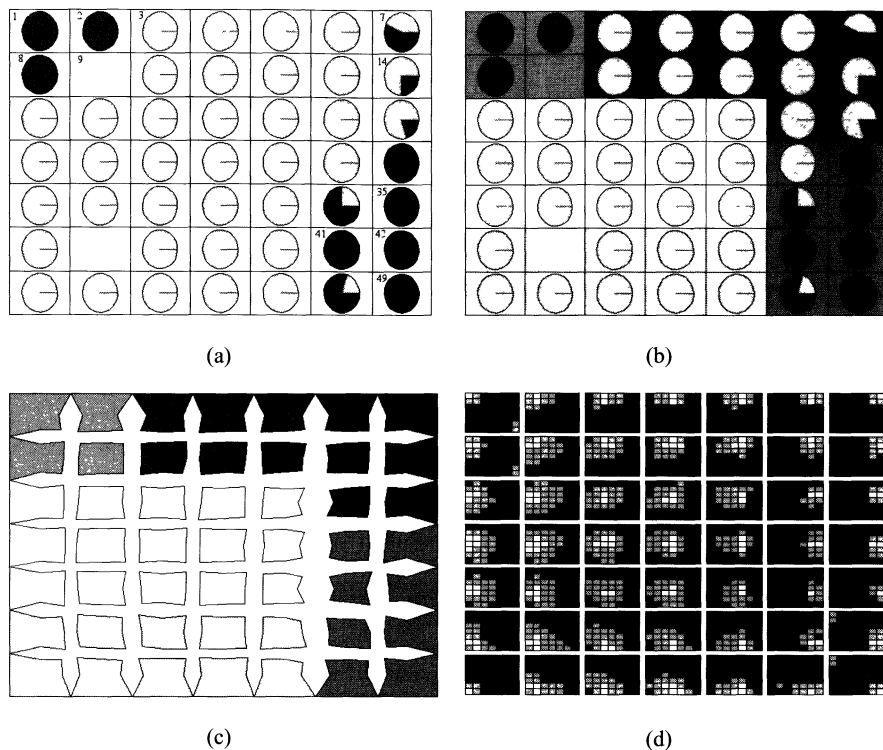


FIGURE 2

Caractérisation de la classification de Kohonen par une variable qualitative endogène ou exogène, unités numérotées par ordre croissant de gauche à droite et de haut en bas :

(a) Fréquence des individus présentant la caractéristique « couleur jaune de la peau », en clair caractéristique absente, en sombre caractéristique présente.

(b) Représentation des macro-classes ajoutée à la carte précédente (solutions du regroupement au niveau 6 d'une classification de type Ward sur les 49 vecteurs codes).



Exemple : 33 % des individus de cette classe ont la peau de couleur jaune.



Niveaux de gris représentant les macro-classes.

Visualisation des distances entre les centres des classes de Kohonen :

(c) Représentation des distances entre les vecteurs codes voisins deux à deux.

(d) Représentation des distances entre tous les vecteurs codes, grilles numérotées par ordre croissant de gauche à droite et de haut en bas.



Niveaux de gris représentant les distances.

haut à gauche et en bas à droite – qui sont tous deux constitués d’individus présentant une peau de « couleur jaune » – sont éloignés ou correspondent à des k -classes proches.

Si la technique des macro-classes augmente notre connaissance de la structure de la carte dans l’espace des données, comme le montre la figure 2b, il est indispensable à ce stade de visualiser la distance entre les vecteurs codes. Une première représentation a été proposée qui visualise la distance entre les vecteurs codes de deux unités voisines (Cottrell-de Bodt 1996). Il s’agit de tracer un bord blanc séparant deux unités voisines d’une épaisseur proportionnelle à la distance entre les vecteurs codes associés. Cette méthode appliquée à l’exemple de la typologie de la peau humaine saine augmente notre connaissance de la structure en montrant quand deux macro-classes séparent une grande population homogène ou quand elles isolent de petites populations atypiques. Mais cette représentation, qui ne prend en compte que les distances entre les k -classes voisines, ne permet pas d’expliquer pourquoi les grandes distances ne coïncident pas forcément avec des changements de macro-classes. Dans l’exemple de la typologie de la peau humaine saine (figure 2c), on peut noter un fossé qui sépare les deux macro-classes (indiquées par le troisième niveau de gris et en gris foncé) à droite du reste de la carte. Mais on peut relever également que la distance entre les vecteurs codes des unités 35 et 41 qui appartiennent à la même macro-classe (troisième niveau de gris) est parmi les plus grandes. De plus, il n’est toujours pas possible de conclure si les individus répartis dans les angles en haut à gauche et en bas à droite sont proches ou non.

Il est donc nécessaire de représenter la distance entre toutes les k -classes, c’est-à-dire U^2 mesures. On présente maintenant une méthode qui consiste à affecter, sur la carte de Kohonen, à chaque couple de k -classes u et u' un niveau de gris qui est d’autant plus foncé que la distance entre leurs vecteurs codes est importante (4 niveaux de gris possibles). Le niveau de gris de la case numérotée u' du $u^{\text{ème}}$ cadre correspond à la distance $d(u, u')$. Cette représentation permet de conserver la structure de voisinage des cartes de Kohonen tout en indiquant la distance entre les centres des k -classes et nous fournit une aide visuelle satisfaisante pour comprendre la structure de l’espace des données. Les grilles étant numérotées par ordre croissant de gauche à droite et de haut en bas, dans l’exemple de la typologie de la peau humaine saine, la figure 2d nous montre sur la première grille en haut à gauche que les unités proches de l’unité 1 (en blanc) sont d’une part les trois unités (2, 8 et 9, premier niveau de gris) qui l’entourent et qui forment avec elle la première macro-classe et d’autre part les unités 42 et 49 dans l’angle en bas à droite (premier niveau de gris). On peut donc maintenant conclure que les deux groupes de k -classes (en haut à gauche et en bas à droite) qui sont chacun caractérisés par la présence « d’une couleur jaune de la peau » – sont voisins, ce qui n’était visible dans aucune des précédentes représentations. Au-delà de la visualisation par la première grille du repliement de la carte dû au rapprochement des angles « haut à gauche » et « bas à droite » de la carte, on peut remarquer que cette représentation des distances met aussi en évidence les phénomènes qui sont déjà apparus avec les précédentes représentations. En particulier, sur les grilles du bord droit, par exemple les grilles 27 et 28, on retrouve l’éloignement du bord droit du reste de la carte signalé dans le précédent paragraphe avec en plus la nuance due à l’exception de l’angle « bas à droite » qui se rapproche de l’angle « haut à gauche » et du bas de la carte, ce qui n’était pas visible précédemment.

5. Représentation d'une classification quelconque sur une carte de Kohonen

Dans cette section, le résultat de l'algorithme de Kohonen est perçu comme un résumé de l'espace des données en U points de l'espace et les différentes cartes présentées dans les sections 2 et 3 comme une visualisation du nuage de points. Il est donc naturel pour résoudre notre objectif initial d'utiliser ces cartes pour représenter les résultats d'une classification (non nécessairement de type Kohonen). Le choix de la distance pour construire la carte étant libre, on considère qu'elle est cohérente avec celle utilisée pour la classification. On construit une variable qualitative Q qui à chaque individu de l'espace des données associe le numéro de sa classe. On peut faire la cartographie de cette variable comme présenté dans la section 3 (figure 2a) et ainsi localiser les différentes classes sur la carte. Les cartes de la section 4 (figures 2b et 2d) permettent de situer les vecteurs codes dans l'espace et donc de connaître la proximité entre les classes. On peut concevoir que plusieurs classifications associées à des distances cohérentes soient chacune adaptées à une base de données et il est donc intéressant de pouvoir comparer les classifications grâce à leur représentation sur la carte. Dans l'exemple de la typologie de la peau humaine saine, on a une première classification de type hiérarchique avec la distance de Ward (en 6 classes) et une deuxième classification obtenue grâce à une méthode de segmentation (en 6 groupes également). Pour ne pas mélanger les deux rôles de l'algorithme de Kohonen que sont la classification et la représentation des données, nous ne considérerons volontairement pas ici la classification de Kohonen comme une troisième approche (à nombre de classes fixé).

Les différentes représentations des deux classifications de la peau humaine saine (figures 3a à 3d) permettent de localiser les 6 classes et les 6 groupes sur la carte alors que la carte des distances (figure 2d) permet de les situer dans l'espace des données. Les figures 3a à 3d visualisent la probabilité empirique pour un individu d'une classe k ou d'un groupe k d'être situé dans la région de l'espace résumée par l'unité i (figures 3c et 3d), et celle d'un individu situé dans l'espace des données autour du vecteur code de l'unité i d'être classé dans la classe k ou le groupe k (figures 3a et 3b). En effet, on note n_{ik} le nombre d'individus de la classe k (respectivement groupe k) associés à l'unité i , n_i le nombre d'individus associés à l'unité i et $n_{.k}$ le nombre d'individus affectés à la classe k (respectivement le groupe k). Sur la figure 2b décrite au premier paragraphe de la section 4, la part du camembert associée au $k^{\text{ème}}$ niveau de gris situé dans l'unité i représente $\frac{n_{ik}}{n_i}$. Sur les figures 3c et 3d, la longueur du bâtonnet de l'histogramme associé au $k^{\text{ème}}$ niveau de gris et situé dans l'unité i représente de la même façon $\frac{n_{ik}}{n_{.k}}$. On rappelle que le niveau de gris du fond indique les unités que l'on peut regrouper au sens des macro-classes. Comme la carte des distances permet de repérer les vecteurs codes les uns par rapport aux autres dans l'espace des données et que l'on vient de voir comment identifier les individus concentrés autour de ces points, on peut situer les classes ou les groupes les uns par rapport aux autres. En particulier, on a vu que le bord de droite s'écarte du reste de la carte excepté l'angle bas à droite. A partir des figures 3a, 3b, 3c, 3d, on constate que le bord droit correspond aux classes 6, 3 et 2 (respectivement les groupes 2, 5 et 6). On conclut donc que les classes 6 et 2 sont loin des classes 1, 4 et 5. De plus, seule la classe 3 présente conjointement dans les angles bas à droite et en haut à gauche

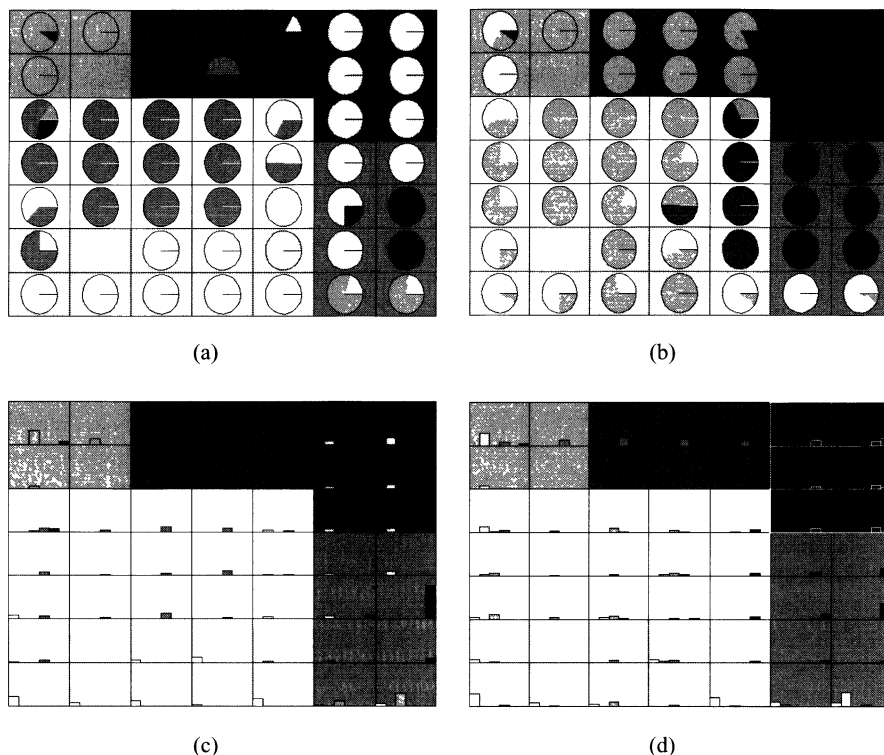


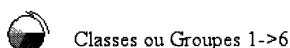
FIGURE 3

La contingence du croisement des k -classes obtenues avec l'algorithme de Kohonen avec celles issues d'une autre classification effectuée sur la même base de données est projetée sur la carte de Kohonen afin d'intégrer la topologie entre les classes.

La contingence est exprimée en pourcentage de l'effectif des k -classes : les classes (a) ou groupes (b) sont indiqués par le niveau de gris des camemberts.

(a) Caractérisation des k -classes par les classes issues de la classification hiérarchique avec la distance de Ward de la même base de données.

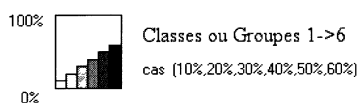
(b) Caractérisation des k -classes par les classes de Kohonen par les groupes issus de la segmentation



La contingence est exprimée en pourcentage de l'effectif de chaque classe (c) ou de chaque groupe (d) : les classes ou groupes sont indiqués par le niveau de gris des bâtonnets.

(c) Répartition des classes issues de la classification hiérarchique avec la distance de Ward dans les k -classes.

(d) Répartition des groupes issus de la segmentation dans les k -classes



se rapproche de ces trois classes (de la classe 1 via l'angle bas droite et des classes 4 et 5 via l'angle haut à gauche). On distingue aussi les individus de la classe 3, qui sont situés dans l'angle haut à gauche et sont donc affectés par la segmentation aux groupes 2 ou 4, et ceux situés dans l'angle bas à droite et donc répartis entre les groupes 1 et 2. En conclusion, la répartition des individus de la classe 3 dans trois groupes 1, 2 et 4 s'explique par la proximité entre les groupes 1 et 2 (angle bas droite) et entre les groupes 2 et 4 (angle haut gauche), en particulier l'hypothèse de la proximité entre les groupes 1 et 4 est écartée. D'une façon plus générale, la différence entre les deux approches correspond au rapprochement des angles haut gauche et bas droite de la carte. La classification l'identifie comme une classe unique alors que la segmentation considère que l'angle haut gauche est le prolongement du bord haut, et respectivement que l'angle bas droite est le prolongement du bord bas.

6. Perspectives

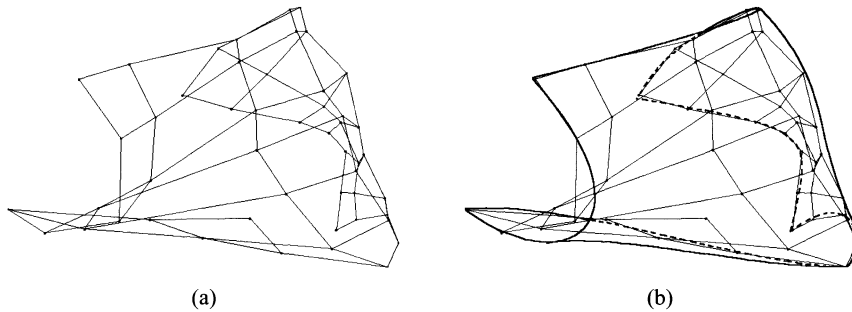


FIGURE 4

Projection sur des plans de la surface générée à partir à la fois des centres de classes de l'algorithme de Kohonen et des liens de voisinages entre eux. Pour une question de lisibilité, seuls les liens entre 4 voisins sont représentés :

- (a) Projection de la surface sur le premier plan principal*
- (b) Surimpression du bord de la carte en trait plein (en pointillé pour les bords cachés par la surface)*

Quand on trace sur un plan factoriel, déterminé à partir de la base de données, les vecteurs codes en reliant ceux dont les k -classes associées sont voisines, on obtient une représentation du type de la figure 4a pour l'exemple de la typologie de la peau humaine. Les liens ainsi constitués déterminent des quadrilatères qui génèrent une surface qui passe par les U points définis par l'algorithme de Kohonen (figure 4b). La représentation du nuage de points par l'ensemble des cartes de Kohonen présentées dans les sections 3, 4 et 5 devient ainsi également une visualisation de cette surface ainsi que de toute surface ajustant le nuage de points et passant par les U points. On a ainsi le moyen de résumer le nuage de points par une surface avec un outil spécifique pour la représenter, en particulier adapté à sa non linéarité.

7. Conclusion

On a présenté une méthode de visualisation de données à partir des cartes de Kohonen qui est une substitution possible à celles de la famille des représentations linéaires quand celles-ci ne sont pas satisfaisantes. Elle apparaît plus adéquate à certaines utilisations comme la représentation graphique des résultats d'une classification pour constituer un support visuel de comparaison de différentes classifications. Elle permet ainsi de confronter différentes approches qui peuvent être contradictoires ou complémentaires. Cette méthode de visualisation de données ouvre aussi la perspective de pouvoir ajuster un nuage de points par des surfaces non linéaires tout en ayant des représentations graphiques de celles-ci.

Remerciements

Les auteurs remercient le professeur E. Tschachler pour ses encouragements et toute l'équipe du C.E.R.I.E.S. pour leur contribution aux données.

Références

- BLAYO F. and DEMARTINES P. (1991), Data analysis : how to compare Kohonen neural networks to other techniques? In *Proceedings of IWANN'91*, pp. 469-476, Springer Verlag, Berlin.
- CHAVENT M., GUINOT C., LECHEVALLIER Y. et TENENHAUS M. (1999), Méthodes divisives de classification et segmentation non supervisée : Recherche d'une typologie de la peau humaine saine. *Revue de Statistique Appliquée*, XLVII, 87-99.
- COTTRELL M. and IBBOU S. (1995), Multiple Correspondence Analysis of a crosstabulations matrix using the Kohonen algorithm. In *Proceedings of ESANN'95*, M. Verleysen (Eds), pp. 27-32, D Facto, Bruxelles.
- COTTRELL M. and de BODT E. (1996), A Kohonen Map Representations to Avoid Misleading Interpretations. In *Proceedings of ESANN'96*, M. Verleysen (Ed.), pp. 103-110, D Facto, Bruxelles.
- COTTRELL M. and ROUSSET P. (1997), A powerful Tool for Analysing and Representing Multidimensional Quantitative and Qualitative Data. In *Proceedings of IWANN'97*, pp. 861-871, Springer Verlag, Berlin.
- COTTRELL M., FORT J.C., PAGÈS G. (1998), Theoretical aspects of the SOM algorithm, *Neuro Computing*, 21, pp. 119-138
- COTTRELL M., GAUBERT P., LETREMY P. and ROUSSET P. (1999), Analyzing and representing multidimensional quantitative and qualitative data : Demographic study of the Rhône valley. The domestic consumption of the Canadian families. In *Kohonen Maps*, E. Oja and S. Kaski (Eds), pp. 1-14, Elsevier, Amsterdam.

- GUINOT C., TENENHAUS M., DUBOURGEAT M., LE FUR I., MORIZOT F. et TSCHACHLER E. (1997), Recherche d'une classification de la peau humaine saine : méthode de classification et méthode de segmentation. *Actes des XXIX^e Journées de Statistique de la SFdS*, pp. 429-432.
- KOHONEN T. (1993), *Self-organization and Associative Memory*. 3^e ed., Springer Verlag, Berlin.
- KOHONEN T. (1995), *Self-Organizing Maps*. Springer Series in Information Sciences Vol. 30, Springer Verlag, Berlin.
- LEBART L., MORINEAU M. et PIRON M. (1995), *Statistique exploratoire multidimensionnelle*. Dunod, Paris.
- LLOYD S.P. (1982), Least squares quantization in PCM. *IEEE Transactions on Information Theory*, IT-28, 2, 129-149.
- ROUSSET P. (1999), Application des algorithmes d'auto-organisation à la classification et à la prévision. Thèse de doctorat, Université Paris I, pp. 41-68.
- WONG M.A. (1982), A hybrid clustering method for identifying high density clusters. *J. Amer. Statist. Assoc.*, 77, 841-847.