CLASSIFICATION D'INDIVIDUS DÉCRITS PAR DES VARIABLES MIXTES STRUCTURÉES EN BLOCS

Ndèye Niang[1], Mory Ouattara[1,2], Julien Brajard[3], Sylvie Thiria[3], Houda Yahi[3], Corinne Mandin[2]

[1]Chaire de Statistique Appliquée & CEDRIC CNAM
292, rue Saint Martin, 75141 Paris Cedex 03, France,
ndeye.niang_keita@cnam.fr,
[2]Centre Scientifique et Technique du Bâtiment
84 Avenue Jean Jaurès
77420 Champs-sur-Marne
mory.ouattara@cstb.fr
corinne.mandin@cstb.fr
[3] UPMC-LOCEAN
4 place Jussieu, BC 100, tout 45-55, 4eme étage
75252 Paris Cedex 05
julien.brajard@locean-ipsl.upmc.fr
sylvie.thiria@locean-ipsl.upmc.fr
houda.yahi@locean-ipsl.upmc.fr

RÉSUMÉ

Nous nous intéressons à la classification d'individus décrits par des variables mixtes structurées en blocs. Nous proposons une méthode de type hiérarchique en deux étapes pour obtenir une typologie des individus. Elle repose sur une combinaison des cartes topologiques mixtes et de la classification hiérarchique ascendante. La méthode proposée permet d'une part le traitement simultané des variables qualitatives et quantitatives et d'autre part elle prend en compte la structuration en blocs des variables. Elle est illustrée sur des données réelles issues de la campagne logements de l'Observatoire de la qualité de l'air intérieur (OQAI).

MOTS-CLÉS: Carte topologique mixte, Kohonen, données mixtes, classification hiérarchique.

SUMMARY

We address the problem of clustering individuals described with mixed variables which are divided in homogeneous blocks. We propose a hierarchical method with two levels to partition the individuals. The method is based on a combination of mixed topological map and ascendant hierarchical clustering. The proposed approach allows to take into account simultaneously qualitative and quantitative variables as well as the variables blocking. A real example illustrates the proposed method.

KEYWORDS: mixed topological map, Kohonen, mixed data, hierarchical clustering.

1 Introduction

Les tableaux de données décrivant un ensemble d'individus à la fois avec des variables quantitatives et qualitatives sont souvent rencontrés dans la pratique. Plusieurs méthodes ont été proposées par différents auteurs pour étudier ces données dites mixtes dans le cadre de l'analyse factorielle. Elles reposent généralement soit sur un codage optimal des modalités des variables qualitatives pour ensuite les traiter comme des variables numériques à travers une analyse en composantes principales (ACP), soit sur une transformation adéquate des variables quantitatives en qualitatives (B. Escofier (1979)) en préalable à une analyse des correspondances multiples (ACM). Saporta (1990) propose de réaliser une ACP, avec une métrique judicieusement choisie, sur le tableau résultant de la juxtaposition des variables quantitatives réduites et des variables qualitatives codées

sous forme disjonctive complète. On peut enfin citer l'analyse factorielle multiple (AFM) proposée par Escofier et Pagès (1998) qui traite le cas où les variables constituent des groupes homogènes quant au type, qualitatif ou quantitatif, des variables. Pagès (2004) regroupe les points de vue d'Escofier et de Saporta dans l'analyse factorielle de données mixtes (AFDM).

Dans le cadre de la classification d'individus, lorsque ces derniers sont décrits par des variables qualitatives, une méthode usuelle consiste à réaliser une ACM au préalable et à effectuer la classification sur les coordonnées factorielles alors quantitatives. En présence de données mixtes, une extension naturelle serait d'appliquer la classification sur coordonnées factorielles aux composantes issues de l'AFDM.

Nous proposons, en préalable à la CAH, d'utiliser les cartes topologiques mixtes ou mixed topological map (MTM) Lebbah (2005) qui sont une extension aux données mixtes des cartes de Kohonen. Cette méthode sera notée MTM-CAH dans la suite.

Dans cette communication, nous nous intéressons à la classification d'un ensemble d'individus décrits par des variables mixtes structurées en blocs. Dans une démarche semblable aux approches d'ACP hiérarchique (Wold (1996)), nous proposons une méthode qui consiste à appliquer MTM-CAH de manière hiérarchique permettant ainsi le traitement simultané de variables qualitatives et quantitatives ainsi que la prise en compte de leur structuration en blocs.
Dans la suite nous présentons les cartes topologiques mixtes MTM en section 2, puis nous décrivons la méthode que nous proposons dans la section 3. La section 4 est consacrée à l'application de la méthode proposée à une base de données réelles issue de la campagne logements de l'Observatoire de la qualité de l'air intérieur organisée entre 2003 et 2005.

2 Classification par cartes topologiques mixtes (MTM-CAH : Mixed Topological Map)

Les cartes topologiques auto-organisées ou self organized map (SOM) font partie de la famille des méthodes neuronales dédiées à la classification non supervisée. Elles sont utilisées pour quantifier et visualiser des données numériques de grandes dimensions dans des espaces de faibles dimensions. L'espace de visualisation usuellement de dimension deux est appelé carte topologique.
De manière générale la méthode suppose l'existence d'une carte discrète C ayant N_c cellules structurées par des graphes non-orientés permettant de définir une distance δ entre deux cellules comme étant la longueur de la plus petite chaîne les reliant.

2.1 L'algorithme de MTM

L'algorithme de kohonen (SOM) a été étendu d'abord au cas de données qualitatives en particulier binaires dans le cadre de l'algorithme Binbatch (Lebbah 2000) et à travers les variantes de l'algorithme KMCA Kohonen Multiple Correspondence Analysis (Cottrell 2004). Ensuite une extension au cas de données mixtes a été proposée à travers l'algorithme MTM que nous présentons ci-dessous.
La méthode MTM traite des jeux de données composés de deux parties : une partie continue regroupant l'ensemble des variables quantitatives et une partie constituée de variables qualitatives mises sous forme disjonctive au préalable.
Chaque observation $\boldsymbol{z_i}$ de la base de données $\boldsymbol{E = \{z_i; i = 1 \dots N\}}$ comporte ainsi une partie numérique $\boldsymbol{z_i^r = (z_{i1} \dots z_{ip})}$ ($\boldsymbol{z_i^r \in \mathrm{R}^p}$) et une partie binaire $\boldsymbol{z_i^b = (z_{i(p+1)} \dots z_{in})}$ $\boldsymbol{z_i^b \in \{0,1\}^{n-p}}$).
On associe alors à chaque cellule c de la carte C un vecteur référent $w_c = (w_c^r, w_c^b)$ où w_c^r dans $\mathbf{R^p}$, désigne la partie réelle du référent c et w_c^b dans $\boldsymbol{\{0,1\}^{n-p}}$ la partie binaire. Ainsi, le vecteur référent a la même structure que les observations $\boldsymbol{z_i}$ de la base initiale.
On note W l'ensemble des vecteurs référents, $\boldsymbol{W^r}$ sa partie réelle et $\boldsymbol{W^b}$ sa partie binaire.
L'algorithme utilisé est une combinaison de l'algorithme de kohonen pour les variables numériques

et de l'algorithme Binbatch pour les variables binaires. Nous décrivons ci dessous les spécificités de la méthode MTM en particulier la fonction de coût.
Cette dernière repose sur une mesure D de dissimilarité entre une observation et un référent spécifique à chaque type de données : la distance euclidienne pour la partie numérique et de la distance de Hamming Ң pour la partie binaire:

$$D(z_i, w_c) = \|z_i - w_c\|^2 = \|{z_i}^r - w_c^r\|^2 + \|{z_i}^b - w_c^b\|^2 = \|{z_i}^r - w_c^r\|^2 + \text{ϐ}\text{Ң}({z_i}^r, w_c^b) \quad \textbf{(1)}$$

où $\text{ϐ} = \frac{p}{n-p}$ est le poids relatif des variables qualitatives par rapport aux variables quantitatives. La fonction de coût $J_{MTM}^T(\chi, w)$ à minimiser est alors décomposée en deux parties $J_{som_r}^T$et $J_{bin_b}^T$ représentant respectivement les fonctions de coût classiques associées à l'algorithme SOM et à l'algorithme Binbatch. On a :

$$\boldsymbol{J_{MTM}^T(\chi, w) = \sum_{z_{i\epsilon E}} \sum_{c\in C} \kappa^T\left(\delta(\chi(z_i), c)\right) * D \quad = \quad J_{som_r}^T(\chi, w_c^r) + J_{bin_b}^T\left(\chi, w_c^b\right)} \qquad \textbf{(2)}$$

χ affecte chaque observation z_i à une cellule de carte et κ^T ($\kappa \geq 0\ et\ \lim_{|x|\to\infty} \kappa(x) = 0$) définit le système de voisinage associé à chaque référent

T le paramètre de voisinage. La minimisation de la fonction de coût passe par une itération en deux phases :

- une phase d'affection qui est une mise à jour de la fonction d'affection χ associée à l'ensemble W fixé, chaque observation z est affectée à un référent c tel que pour tout z_i :

$$\boldsymbol{\chi(z_i) = argmin_c(\|z_i - w_c\|^2)} \qquad \textbf{(3)}$$

- La phase d'optimisation consiste à choisir pour χ fixé le système de référents W qui minimise les fonctions $\boldsymbol{J_{som_r}^T(\chi, w_c^r)}$ et $\boldsymbol{J_{bin_b}^T\left(\chi, w_c^b\right)}$, conduisant aux expressions suivantes de calcul des référents : $w_c^r = \frac{\sum_{z_{i\epsilon E}} \kappa(\delta(\chi(z_i), r)) {z_i}^r}{\sum_{z_{i\epsilon E}} \kappa(\delta(\chi(z_i), r))}$ pour partie réelle et $W_c^{bk} = \begin{cases} 0\ si\ \sum_{z_{i\epsilon A}} \kappa\left(\delta(\chi(z_i), r)\right)(1 - {z_i}^{bk}) > \sum_{z_{i\epsilon A}} \kappa\left(\delta(\chi(z_i), r)\right) {z_i}^{bk} \\ 1\ sinon \end{cases}$ pour les coordonnées de la partie binaire.

L'algorithme de MTM se déroule alors en initialisant à t=0 les p référents initiaux, la structure et la taille de la carte, le nombre d'itérations et le paramètre de voisinage T puis en appliquant les phases d'affectation et d'optimisation de la fonction de coût (2).
On répète l'étape itérative jusqu'à ce que l'on atteigne un nombre fixé N_{iter} d'itérations ou jusqu'à la convergence de la carte. Le nombre de neurones est alors choisi de façon empirique après plusieurs expériences de façon à capter la structure des données. La carte finale est retenue suivant deux critères : un indice de similarité (par exemple l'indice de Jaccard) qui permet d'apprécier l'homogénéité des observations captés par un référent et une contrainte de voisinage telle que deux neurones proches sur la carte se ressemblent dans l'espace des données.
A la fin de l'algorithme, chaque référent aura capté un groupe d'observations fournissant ainsi une partition de l'ensemble des individus.

2.2 Réduction du nombre de classe par MTM

Généralement le nombre de classes de la partition résultante de MTM est trop grand pour permettre

une interprétation aisée, par exemple dans nos applications le nombre de référents est de l'ordre de 100.
Comme dans Yacoub (2001), nous proposons donc, à la suite des cartes topologiques MTM d'effectuer une classification ascendante hiérarchique sur les référents permettant d'obtenir un regroupement des observations en un nombre restreint de classes.
Rappelons que les référents sont dans le même espace que les données. La CAH doit donc être réalisée avec la même distance que celle utilisée par l'algorithme MTM. On choisit donc la distance *D* définie en (1) comme critère de classification dans la CAH.

3 Méthode proposée

Nous nous intéressons au cas de données mixtes structurées en blocs (figure 1-a). La structuration en blocs est induite par un regroupement des variables selon des critères permettant de caractériser les observations par rapport à des thèmes spécifiques et prédéfinis. Dans le cas particulier de notre application il s'agit de la structure technique des logements, la structure des ménages et les habitudes de vie des habitants.

	Bloc 1	
1 . . n	$X_{r11}...X_{r1p1}$ P_1 variables réelles	$X_{b11}...X_{bc1}$ Q_1 variables qualitatives

...

	Bloc P	
1. . . n	$X_{r1}...X_{rp}$ P_p variables réelles	$X_{r1}...X_{rk}$ Q_p variables qualitatives

Figure 1-a

Pour prendre en compte cette structuration en blocs, nous proposons une méthode hiérarchique à deux niveaux comme dans l'ACP hiérarchique de Wold (1996). Le premier, appelé niveau inférieur consiste à appliquer MTM-CAH sur chacun des blocs. On obtient ainsi des typologies des individus propres aux variables de chaque groupe. Il est alors possible, à travers l'interprétation des classes, de dégager des caractéristiques des groupes homogènes d'individus selon les variables de chaque bloc. Le deuxième niveau dit supérieur, consiste à appliquer une seconde fois MTM-CAH aux variables déduites des typologies issues du niveau inférieur afin d'obtenir une unique classification finale. Cette seconde étape permet donc de prendre en compte simultanément les informations issues de chaque bloc de variables. Les variables résultantes du premier niveau sont ici supposées avoir la même importance, on leur associe un poids identique à la différence de l'ACP hiérarchique. Notre démarche est résumée dans la figure 1 b.

4 Application

Dans cette section, après avoir décrit les données, nous présenterons les résultats de l'application de la méthode précédemment présentée. Ces derniers sont ensuite comparés aux résultats obtenus par la démarche plus classique qui consiste à appliquer l'ACM en préalable à la CAH.

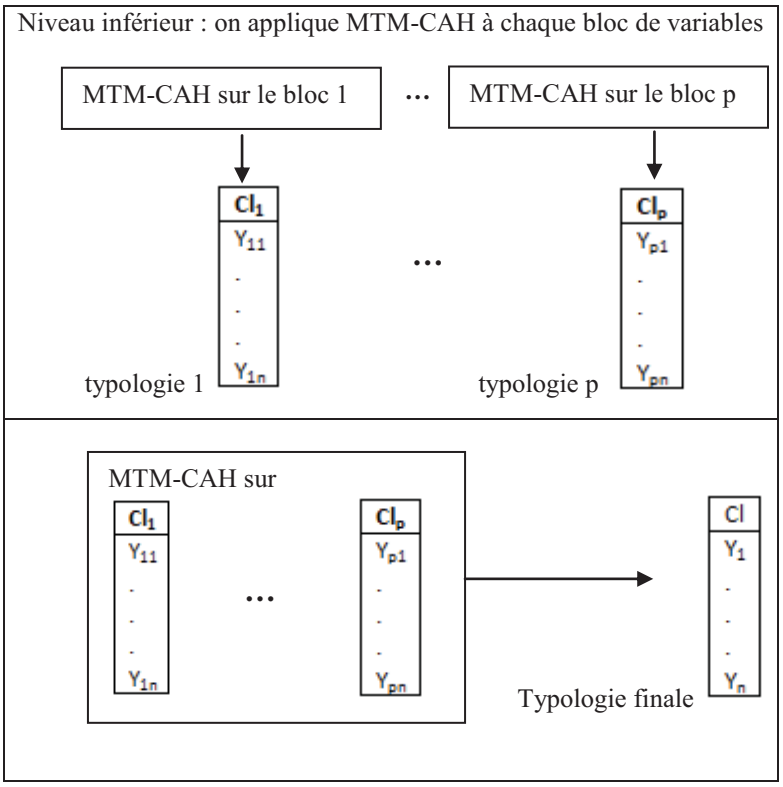

Figure 1-b: Niveau supérieur : on applique MTM-CAH sur les résultats du niveau inférieur

4.1 Le jeu de données

L'Observatoire de la Qualité de l'Air Intérieur (OQAI) a réalisé entre 2003 et 2005 une campagne nationale de mesure de la pollution intérieure sur un échantillon de 567 logements représentatifs du parc des 24 millions de résidences principales de France métropolitaine continentale (Kirchner (2007)). Cette campagne visait à dresser un état de la pollution de l'air dans l'habitat afin de donner les éléments utiles pour l'estimation de l'exposition des populations, la quantification et la hiérarchisation des risques sanitaires associés, ainsi que l'identification des facteurs prédictifs de la qualité de l'air intérieur.
Les données récoltées regroupent plus de 125 variables et 567 observations (logements). Elles sont divisées en trois blocs suivant les critères permettant de les caractériser par rapport à la structure des logements, des ménages (occupants) et celle des habitudes de vie des habitants. Chacun des trois blocs de variables est composé d'un groupe de variables quantitatives et d'un groupe de variables qualitatives.

4.2 Résultats

L'application de MTM-CAH sur les blocs associés aux trois critères permet d'obtenir des typologies dites de logements, de ménages et d'habitudes regroupant les individus en classes homogènes et interprétables. On obtient par exemple une classe de la typologie « logements » correspondant aux maisons individuelles récentes « tout en un » plutôt petites, propriété des occupants, avec un jardin. Des Garages attenant et communicant à la partie habitée sont présents presque systématiquement dans cette classe. Ces maisons ne disposent pas de cheminée. Elles sont équipées de système de chauffage.
Comme proposée dans la deuxième étape de la méthode décrite dans la section 3, nous avons appliqué MTM-CAH aux variables issues des trois typologies. Sur la figure 2, on constate que la contrainte de voisinage a surtout structuré la typologie des logements mais certaines relations de

voisinage sont aussi visibles sur la typologie des habitudes. Certains référents ne comportent aucune observation (ils sont dits blancs) et permettent de séparer l'espace entre des groupes de données très différents tout en conservant la contrainte de voisinage. L'indice de Jaccard évaluant l'homogénéité des observations captées par les référents de la carte retenue ici vaut 0.80.

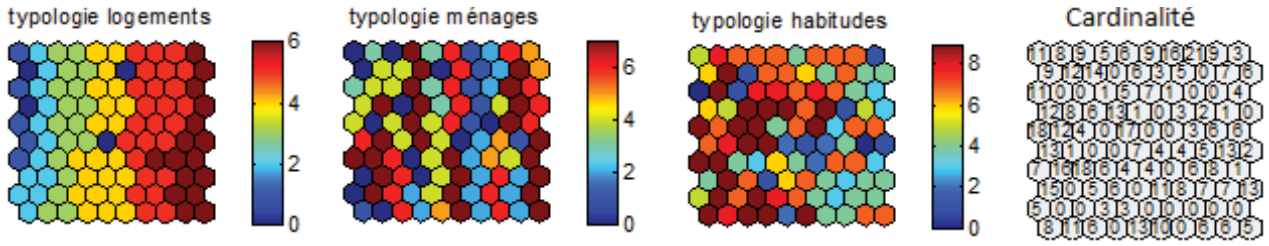

Figure 2: representation de la carte MTM de la valeurs des référents pour les typologies; les codes couleurs symbolisent les classes des typologies.

Finalement, on a effectué une CAH en utilisant la stratégie du lien moyen. On obtient une partition en 8 classes permettant de caractériser les individus à la fois par rapport aux ménages, habitudes et logements mettant alors en évidence des liaisons des variables inter-blocs.
Les variables issues de typologies étant toutes qualitatives, dans la deuxième phase, nous avons aussi réalisé une ACM suivie d'une CAH. L'indice de Rand utilisé pour comparer les deux partitions vaut 0.80 traduisant une forte ressemblance quant au contenu des classes. Une étude comparative des interprétations des classes permettrait d'affiner ce résultat.

5 Conclusion

Dans le cadre de la classification d'individus décrits par des variables mixtes structurées en blocs homogènes, nous avons proposé une méthode qui permet le traitement simultané des variables qualitatives et quantitatives et qui prend en compte la structuration en blocs des variables.

La méthode semblable à celle de l'ACP hiérarchique pourrait cependant être améliorée en y intégrant un système de pondération des variables résultantes de la première étape.

Des évaluations plus formelles des performances de la méthode sont nécessaires ainsi que des études comparatives avec des approches basées sur d'autres méthodes proposées dans la littérature telles que l'analyse factorielle de données mixtes.

Bibliographie

[1] Cottrell M,Smaïl I,Patrick L (2004) *SOM-based algorithms for qualitative variables*

[2] Escofier B (1979) *traitement simultané de variables quantitative et qualitative en analyse factorielle,* 4(2) 137-146.

[3] Escofier B et Pages J. (1998) *Analyses factorielle simple et multiples,* 3e ed, Dunod.

[4] Pages J (2004), *analyse factorielle de données mixtes,* revue de statistique appliquée*, tome 52 n°24, P93-111*

[5] Lebbah M, Badran F, Thiria S et Chazotte A (2005) *Mixed Map Topological*, EGC'05, Paris

[6] Yacoub M, Niang N, Badran F, S. Thiria (2001) *A New Hierarchical Clustering Method using Topological Map, ASMDA2001 : 10th Int. Symp. On applied stochastic models and data analysis ,1023 1028*

[7] Saporta G (1990) *simultaneous analysis of qualitative and quantitative data atti della XXXV riunione scientifica ; società itialana di statistica, 6-72*.

[8] Kirchner S, Arènes J-F, Cochet C, Derbez M et al (2007) *État de la qualité de l'air dans les logements français. Environnement, Risques & Santé 2007* ; 6(4) : 259-269.

[9] Wold S, Kettaneh N (1996) *Hierarichal Multiblock PLS and PC models interpretation and as an aternative to variable selection*