2S-SOM : une méthode de soft-subspace clustering pour données multi-blocs basée sur les cartes topologiques auto-organisées

Mory Ouattara[*,**] Ndèye Niang[*] Fouad Badran[*] Corinne Mandin[**]

[*]Statistique Appliquée, CNAM 292, rue Saint Martin, 75141 Paris Cedex 03, France,
n-deye.niang_keita@cnam.fr,
fouad.badran@cnam.fr,
[**]Centre Scientifique et Technique du Bâtiment
84 Avenue Jean Jaurès, 77420 Champs-sur-Marne
mory.ouattara@cstb.fr,
corinne.mandin@cstb.fr,

Résumé. Nous proposons une méthode de soft subspace clustering basée sur les cartes topologiques pour la classification d'individus décrits par des variables structurées en blocs homogènes. L'algorithme nommé Soft Subspace SOM (2S-SOM) consiste à optimiser la fonction de coût de SOM modifiée en introduisant des poids adaptatifs sur les blocs et sur les variables de chaque bloc. Cette double pondération permet de distinguer les blocs les plus importants prenant ainsi en compte la structuration en blocs, et d'identifier pour chaque bloc les variables les plus informatives pour les classes. La méthode permet alors de déterminer simultanément les groupes d'individus et leurs sous espaces caractéristiques optimaux. La méthode est illustrée sur des données réelles issues des bases de l'UCI repository of machine learning et sur des données simulées.

1 Introduction

Les méthodes de classification non-supervisées (ou clustering) permettent d'explorer des données non-labélisées dans le but de trouver des groupes d'observations homogènes et bien séparés. Les récentes avancées technologiques en capacité de stockage d'informations d'une part, et la multiplication des sources d'informations d'autre part, contribuent à la mise en place de bases de données complexes et de grande dimension. Dans des domaines tels que la génétique, la finance, le traitement de données textuelles et les études environnementales, par exemple, on rencontre des données de grande dimensions. Ce qui conduit à avoir plusieurs blocs de variables caractérisant chacune une vue particulière sur les données, on parle de données multi-vues ou multi-blocs. C'est notamment le cas des études environnementales sur la pollution de l'air intérieur où les vues sont associées à des thématiques précises : concentrations des polluants de l'air intérieur, informations collectées sur la santé des occupants et sur l'aménagement des environnements intérieurs (Kirchner et al., 2011). Par ailleurs, l'usage de capteurs est souvent nécessaire pour mesurer les concentrations des polluants et la possible

défaillance de ces capteurs peut alors engendrer des données manquantes ou aberrantes.
Considérer les données multi-blocs en classification comme une seule entraîne généralement une perte du pouvoir discriminant de la notion de distance au fur et à mesure que la dimension augmente. Autrement dit, les observations sont pratiquement toute équidistantes les unes par rapport aux autres (Parsons et al., 2004) quand l'espace devient très grand et les mesures classiques de distance ne permettent généralement pas de distinguer les points proches des points éloignés. En outre, en l'absence d'une structure globale de corrélation entre les variables (la présence possible de variables souvent distribuées uniformément), la similarité entre deux observations est souvent portée par un nombre limité de variables. Les classes sont donc recherchées dans des sous-espaces de l'espace initial, on parle alors de *subpace clustering*. Les méthodes de type subspace clustering reposent sur la recherche de sous espaces de l'espace initial permettant une meilleure détection et interprétation des groupes d'individus (Agrawal et al., 1998; Kriegel et al., 2009). Plus récemment, des approches basées sur la définition d'une pondération des variables ou des blocs permettent de prendre en compte en plus de la grande dimension, la structure multi-blocs (Huang et Ng, 2005; Jing et al., 2007; Chen et al., 2012). Cependant, ces méthodes ne sont pas toujours adaptées à la présence de données manquantes ou aberrantes et à la visualisation des classes dans des espaces de faible dimension.
Nous proposons, 2S-SOM, un algorithme de type subspace clustering basé sur les cartes auto-organisées SOM (Kohonen, 1998). 2S-SOM permet de tenir compte de la dimension élevée des données, de la structure multi-bloc, de la présence de données manquantes ou aberrantes et de faciliter la comparaison des classes et la visualisation des données. La méthode est basée sur une version modifiée de la fonction de coût de SOM en introduisant des poids adaptatifs sur les blocs et sur les variables. L'idée de base consiste à rechercher itérativement une partition des observations et à déterminer pour chaque classe des variables et des blocs spécifiques.
La méthode est présentée dans la section 2 après un bref rappel sur SOM. La section 3 présente les propriétés de 2S-SOM. La méthode est illustrée sur des données réelles en section 4. La section 5 est consacrée à une discussion pour conclure.

2 La méthode 2S-SOM

2.1 Notations et rappels

Nous disposons de N observations z_i décrites par m variables divisées en K blocs. On note :

- $\mathcal{Z}$ la matrice de N observations $z_i \in \mathbb{R}^m$ avec $i = 1, \ldots, N$.
- $\mathcal{V} = \{v_l, l = 1, \ldots, m\}$ l'ensemble des variables divisé en K blocs de p_k variables tels que $p_1 + \ldots + p_k + \ldots + p_K = m$.
- α est une matrice $N_c \times K$ où N_c désigne le nombre de classes c dans $\mathcal{Z}$, α_{ck} est le poids du bloc k dans la classe c.
- $\beta = [\beta_1, \ldots, \beta_K]$ est une matrice $N_c \times m$ où β_k est une matrice de dimension $N_c \times p_k$ définit les poids $\beta_{ckj} (j = 1, \ldots, p_k)$ sur les variables du bloc k pour chaque classe.

On cherche donc une partition de $\mathcal{Z}$ en N_c classes.
Les cartes topologiques auto-organisées sont utilisées pour quantifier et visualiser des données numériques de grande dimension dans un espace de faible dimension, généralement 1 ou 2 dimensions, appelé carte topologique. De manière générale, la méthode suppose l'existence

d'une carte discrète $\mathcal{C}$ ayant N_c cellules c structurées par des graphes non-orientés permettant de définir a priori une distance σ entre deux cellules. Dans la suite, nous utiliserons indifféremment les termes cellule ou classe. Chaque cellule de la carte est représentée par un vecteur référent ou prototype w_c synthétisant l'information de la cellule. L'algorithme SOM consiste à optimiser de manière itérative en deux phases la fonction de coût $\mathcal{J}^T_{SOM}$ définie en (1).

$$\mathcal{J}^T_{SOM}(\mathcal{Z},\mathcal{W}) = \sum_{z_i\in\mathcal{Z}}\sum_{c\in C}\mathcal{K}^T(\sigma(\mathcal{X}(z_i),c))\sum_{l=1}^{m}(z_{il}-\omega_{cl})^2 \quad (1)$$

Dans cette expression, $\mathcal{X}(z_i) = \underset{c\in\mathcal{C}}{argmin}(\sum_{r\in\mathcal{C}}\mathcal{K}^T(\sigma(r,c))||z_i-w_c||^2)$ représente une fonction d'affectation des observations z_i à la cellule c dont le vecteur référent est le plus proche, $\mathcal{W}$ est l'ensemble des vecteurs référents ω_c des cellules c. $\mathcal{K}^T$ et le paramètre T associé définissent respectivement une fonction décroissante de contrainte de voisinage définie entre deux cellules c et r de la carte et la taille du voisinage d'une cellule.
L'optimisation de la fonction objectif J^T_{SOM} conduit à l'affectation de chaque individu à une cellule c de la carte représentée par son vecteur référent qui se définit dans la version batch de SOM (Kohonen, 1999) comme la moyenne pondérée des observations de la classe c (cf. 2).

$$\omega^T_c = \frac{\sum_{i=1}^{n}\mathcal{K}^T(\sigma(\mathcal{X}(z_i),c))z_i}{\sum_{i=1}^{n}\mathcal{K}^T(\sigma(\mathcal{X}(z_i),c))} \quad (2)$$

L'une des propriétés les plus importantes de l'algorithme SOM est de rendre possible la comparaison des regroupements réalisés directement à partir des données à l'aide de contrainte de voisinage. En d'autres termes, l'algorithme SOM reproduit la structure de proximité des observations dans l'espace des données sur une carte topologique. De plus, en présence de données de grande dimension, la méthode permet de visualiser les observations dans des espaces de faible dimension, de surmonter la présence de données aberrantes qui sont isolées dans des cellules elles mêmes isolées sur la carte (Kaski, 1997; Kohonen, 1999; Cottrell et al., 2003).
Dans cet article, nous proposons une méthode de classification, nommée 2S-SOM, qui s'intéresse à un même ensemble d'individus décrit par plusieurs blocs de données de grande dimension définis selon des thématiques spécifiques. L'idée de base consiste à conserver les propriétés initiales de SOM et à déterminer le sous-espace caractéristique de chaque classe de la carte topologique.

2.2 2S-SOM

2S-SOM est une extension de l'algorithme de type subspace clustering FGKM basé sur la méthode des K-moyennes (Chen et al., 2012). Il repose sur une modification de la fonction de coût de SOM en introduisant un double système de poids α_{ck} ($k = 1,\ldots,K$) et β_{ckj} ($j = 1,\ldots,p_k$) définis respectivement sur les blocs et sur les variables pour chaque cellule c de la carte $\mathcal{C}$. La classification est donc obtenue par optimisation de la fonction objectif J_{2S-SOM} définie en (3).

$$\mathcal{J}^T_{2S-SOM}(\mathcal{X},\mathcal{W},\alpha,\beta) = \sum_{c\in C}\left(\sum_{k=1}^{K}\left(\sum_{z_i\in Z}\alpha_{ck}\mathcal{K}^T(\sigma(\mathcal{X}(z_i),c))d_{\beta_{ck}} + J_{ck}\right) + I_c\right) \quad (3)$$

avec $d_{\beta_{ck}} = \sum_{j=1}^{p_k} \beta_{ckj}(z_{ikj} - \omega_{ckj})^2$ et sous les contraintes :

$$\begin{cases} \sum_{j=1}^{p_k} \beta_{ckj} = 1,\ \beta_{ckj} \in [0,1], \forall\, c \in \mathcal{C}, \forall k \\ \\ \sum_{k=1}^{K} \alpha_{ck} = 1,\ \alpha_{ck} \in [0,1],\ \forall\, c \in \mathcal{C} \end{cases}$$

$I_c = \lambda \sum_{k=1}^{K} \alpha_{ck} log(\alpha_{ck})$ et $J_{ck} = \eta \sum_{j=1}^{p_k} \beta_{ckj} log(\beta_{ckj})$ représentent les entropies négatives pondérées et associées aux vecteurs poids relatifs aux blocs et aux variables pour une cellule c. Ces termes permettent d'ajuster, selon les paramètres λ et η, les contributions relatives apportées par les variables et les blocs dans la classification. Cela sera détaillé dans la section 3.
L'optimisation de la fonction de coût $\mathcal{J}_{2S-SOM}$ s'effectue de façon alternée en quatre étapes : Les deux premières phases d'affectation des observations aux classes et d'actualisation des vecteurs référents sont identiques à celles de SOM. Dans ces deux premières étapes les valeurs des poids sont supposées connues et fixées à leur valeur courante. On a alors :

– Étape 1 : Les référents $\mathcal{W}$ sont connus et fixés, les observations sont affectées aux cellules en respectant l'équation (4) :

$$c_g(z_i) = \mathcal{X}(z_i) = \underset{c \in \mathcal{C}}{argmin} \left(\sum_{r \in \mathcal{C}} \mathcal{K}^T(\sigma(r,c)) \left(\sum_{k=1}^{K} \alpha_{ck} d_{\beta_{ck}} \right) \right) \tag{4}$$

– Étape 2 : Actualisation des centres de classe à l'aide de (5)

$$\omega_{c_g}^T = \frac{\sum_{z_i \in Z} \mathcal{K}^T(\sigma(X(z_i), c_g)) z_i}{\sum_{z_i \in Z} \mathcal{K}^T(\sigma(\mathcal{X}(z_i), c_g))} \tag{5}$$

Les étapes 3 et 4 suivantes sont similaires à celles de FGKM proposée par (Chen et al., 2012)

– Etape 3 : on démontre que si les paramètres $\mathcal{X} = \widehat{\mathcal{X}}, \omega = \widehat{\omega}$ et $\beta = \widehat{\beta}$ sont connus et fixés à leurs valeurs courantes alors $\forall\ \lambda > 0$, $\mathcal{J}_{2S-SOM}^T(\widehat{\mathcal{X}}, \widehat{\omega}, \alpha, \widehat{\beta})$ atteint son minimum pour une cellule c et pour un bloc k en

$$\alpha_{ck} = \frac{exp(\frac{-\Psi_{ck}}{\lambda})}{\sum_{k=1}^{K} exp(\frac{-\Psi_{ck}}{\lambda})} \tag{6}$$

avec

$$\Psi_{ck} = \sum_{z_i \in \mathcal{Z}} \mathcal{K}^T(\sigma(\mathcal{X}(z_i), c)) d_{\beta_{ck}} \tag{7}$$

qui se décompose dans (8) comme la somme pondérée de la distance des observations z_i des cellules r du voisinage T de la cellule c et en la somme pondérée de la distance des observations $z_i \in c$ au centre de classe de c.

$$\Psi_{ck} = \sum_{z_i \in r, r \neq c} \mathcal{K}^T(r,c) d_{\beta_{ck}} + \mathcal{K}^T(c,c) \sum_{z_i \in c} d_{\beta_{ck}} \tag{8}$$

Le premier terme de (8) est proportionnel à l'inertie des observations appartenant aux cellules r du voisinage T de la cellule c par rapport au centre de classe de la cellule c et le second terme est proportionnel à l'inertie des observations z_i de la cellule c. Finalement, le poids d'un bloc sera donc d'autant plus important que ce bloc minimise simultanément l'inertie des observations appartenant à la classe et l'inertie des observations appartenant au voisinage T de la cellule c.

– Etape 4 : de manière identique, si les paramètres $\mathcal{X} = \widehat{\mathcal{X}}$, $\omega = \widehat{\omega}$ et $\alpha = \widehat{\alpha}$ sont connus et fixés à leurs valeurs courantes alors $\forall\ \eta > 0$, $\mathcal{J}^T_{2S-SOM}(\widehat{\mathcal{X}}, \widehat{\omega}, \widehat{\alpha}, \beta)$ atteint son minimum pour une cellule c et pour un bloc k en

$$\beta_{ckj} = \frac{\exp(\frac{-\Phi_{ckj}}{\eta})}{\sum_{j=1}^{p_k} \exp(\frac{-\Phi_{ckj}}{\eta})} \tag{9}$$

avec

$$\Phi_{ckj} = \sum_{z_i \in \mathcal{Z}} \alpha_{ck} \mathcal{K}^T(\sigma(\mathcal{X}(z_i), c))(z_{ikj} - \omega_{ckj})^2 \tag{10}$$

qui se décompose dans (11), pour la variable j, en la somme pondérée de la distance entre les observations z_i appartenant aux cellules r du voisinage T de c et en la somme pondérée de la distance des observations z_i au référent de la cellule c.

$$\Phi_{ckj} = \sum_{z_i \in r, r \neq c} \alpha_{ck} \mathcal{K}^T(r,c)(z_{ikj} - \omega_{ckj})^2 + \mathcal{K}^T(c,c) \sum_{z_i \in c} \alpha_{ck}(z_{ikj} - \omega_{ckj})^2 \tag{11}$$

Le poids d'une variable sera donc d'autant plus important qu'elle minimise simultanément la variance des observations appartenant à la classe c et la somme des distances entre les observations appartenant aux cellules r du voisinage T de la cellule c et le référent w_c de la cellule. Les coefficients de pondération α_{ck} et β_{ckj} définis par 2S-SOM indiquent respectivement l'importance relative des blocs et des variables dans les classes. Ainsi, plus le poids d'un bloc k ou d'une variable v_j est important, plus le bloc ou la variable contribue à la définition de la classe au sens où elle permet de réduire la variabilité des observations dans la cellule et dans son voisinage proche. Finalement, à la convergence, 2S-SOM fournit d'une part une carte topologique permettant de visualiser les données et d'autre part des systèmes de poids pour les classes de la classification.

Algorithme 2.1 *L'algorithme* $2S - SOM$

1. ***Initialisation :*** *choisir la dimension de la carte, le voisinage initial T_o et final T_f des cellules, définir l'ensemble des centres de classe $\mathcal{W}^0$, initialiser les poids α^0_{ck} sur les blocs et les poids β^0_{ckj} sur les variables et le couple de paramètres λ et η.*
2. ***Affectation :*** *utiliser la formule (4) pour affecter chaque observation à sa cellule d'appartenance.*

3. ***Actualisation des centres de classe :*** *utiliser la formule (5) pour actualiser les référents des cellules.*
4. ***Actualisation des poids sur les blocs :*** *utiliser les formules (6 et 7) pour actualiser les poids sur les blocs.*
5. ***Actualisation des poids sur les variables :*** *utiliser les formules (9 et 10) pour actualiser les poids sur les variables.*
6. ***Répéter*** *les étapes 2 à 5 jusqu'à la convergence de l'algorithme vers un minimum.*

En vue de faciliter l'interprétation des classes, si la taille de la carte conduit à un trop grand nombre de cellules, il est possible d'appliquer un algorithme de classification ascendante hiérarchique (CAH) sous contrainte de voisinage sur la matrice composée des vecteurs référents pour réduire ce grand nombre de cellules en un nombre restreint de classes (Gordon, 1996; Vesanto et al., 2000). La contrainte de voisinage dans la CAH permet alors de conserver la topologie des observations fournie par la carte 2S-SOM. Dans le cas des évaluations présentées dans la section 4, les classes finales sont obtenues par application d'une CAH sous contraintes utilisant la stratégie d'agrégation de ward.

3 Propriétés de 2S-SOM

Dans cette section, nous étudions les propriétés de conservation topologique de 2S-SOM et l'influence des poids α et β dans la classification en fonction des paramètres λ, η et du paramètre de voisinage T. Remarquons que si λ et η sont très grands $\alpha_{ck} \approx \frac{1}{K}$ et $\beta_{ckj} \approx \frac{1}{p_k}$. La fonction objectif $\mathcal{J}^T_{2S-SOM}$ peut se décomposer en (12) faisant apparaître les termes de conservation de la topologie des observations et de quantification vectorielle de 2S-SOM :

$$\begin{aligned}\mathcal{J}^T_{2S-SOM}(\mathcal{X},\mathcal{W},\alpha,\beta) &= \sum_{c\in C}\left(\sum_{k=1}^{K}\left(\sum_{r\in\mathcal{C}}\sum_{z_i\in r}\alpha_{ck}\mathcal{K}^T(\sigma(r,c))d_{\beta_{ck}} + J_{ck}\right) + I_c\right) \\ &= \mathcal{K}^T(\sigma(c,c))\sum_{c\in C}\left(\sum_{k=1}^{K}\left(\sum_{z_i\in c}\alpha_{ck}d_{\beta_{ck}} + J_{ck}\right) + I_c\right) + \\ &\qquad \sum_{c\in C}\left(\sum_{k=1}^{K}\left(\sum_{r\neq c}\sum_{z_i\in r}\alpha_{ck}\mathcal{K}^T(\sigma(r,c))d_{\beta_{ck}}\right)\right) \quad (12)\end{aligned}$$

1. Le premier terme correspond à la fonction objectif proposée par (Chen et al., 2012) dans FGKM pondérée par $\mathcal{K}^T(\sigma(c,c)) = \mathcal{K}^T(0)$. Son importance relative dans 2S-SOM dépend alors de T ; plus T est petit plus ce terme prend de l'importance dans la minimisation, dans ce cas 2S-SOM est équivalent à FGKM. De plus, lorsque les paramètres λ et η sont très grands, les poids α_{ck} et β_{ckj} deviennent constant alors, 2S-SOM est équivalent à l'algorithme des K-Means.
2. Le deuxième terme introduit la contrainte de conservation topologique. Ce terme montre que si deux cellules sont proches sur la carte $\mathcal{C}$ alors $\mathcal{K}^T$ est grand car $\sigma(c,r)$ est petit ; la minimisation de ce terme rapproche les deux cellules r et c. Ainsi, la proximité sur la

carte traduit donc une proximité dans l'espace des observations. De plus, lorsque $\lambda \to \infty$ et $\eta \to \infty$, les blocs sont équipondérés de même que les variables. Ainsi, 2S-SOM est équivalent à SOM si T est grand.

Lorsque $\lambda \to \infty$ et η fixé les poids α_{ck} associés aux blocs étant tous égaux à $\frac{1}{K}$, alors, seuls les poids β_{ckj} des variables des blocs définissent les cellules. En considérant que le sous espace associé à la classe est donné par les variables ayant les plus forts poids, 2S-SOM peut être vu comme un algorithme de soft subspace clustering.
Lorsque $\eta \to \infty$ et λ fixé les poids des variables d'un bloc sont identiques à $\frac{1}{p_k}$. Dans ce cas seuls les blocs sont pénalisés selon leur capacité à définir les cellules. Dans ce contexte, 2S-SOM permet alors de déterminer pour chaque cellule les blocs qui lui sont spécifiques.

4 Applications

4.1 Données

La méthode proposée est illustrée sur 3 bases de données étiquetées disponibles sur le site de UCI Machine learning Repository et sur 3 bases de données simulées. Il sera ainsi possible d'évaluer les performances de l'algorithme à l'aide d'indices externes de performances tels que la précision, le rappel, la F-mesure et la pureté.
La base "Image Segmentation" (IS) contient 2310 observations et 19 variables décrivant les pixels de 7 images. Chaque observation représente un point d'une image décrite par deux blocs de 9 et 10 variables caractérisant le contraste de couleur de ce point sur l'image.
Le jeu de données CT contient 2126 cardiotocographies fœtales décrites par 21 variables regroupées en 3 blocs. Le bloc 1 contient 7 variables liées à la fréquence cardiaque d'un fœtus. Le bloc 2 contient 4 variables décrivant la variabilité de rythme cardiaque et le bloc 3 est composé de 10 variables définissant des histogrammes de la cardiotocographie (CT) du fœtus. Ces 2126 observations sont divisées en 10 classes.
Le troisième jeu de données "Dutch utility maps" (DMU) est composé de 2000 observations correspondant à l'écriture manuelle des 10 chiffres de la numération mathématique (0 à 9). La représentation sous forme d'image de ces 2000 observations est décrite par 649 variables structurées en 6 blocs contenant respectivement 76, 216, 64, 240, 47 et 6.
Les données simulées D1 et D2 contiennent chacune 400 observations divisées en 4 classes de 100 observations décrites par 4 blocs de variables. La table D1 contient 100 variables réparties en 4 blocs de 20 variables et la table D2 contient 4 blocs de 5 variables. La table D3 contient 400 observations et 25 variables, elle est obtenue en rajoutant à la table D2 un bloc de 5 variables de bruit et 5% d'observations aberrantes. Le tableau 1 présente les caractéristiques des données en terme de structuration en blocs et de répartition des variables de bruit dans les blocs ainsi que les paramètres des cartes topologiques retenus. La figure 1 représente les classes des données simulées dans le plan factoriel d'une ACP sur les 3 tables.

Pour chaque jeu de données 2S-SOM a été appliqué à travers plusieurs initialisations des paramètres de l'algorithme : choix des centres initiaux, dimensions de la carte, taille du voisinage, nombre d'itérations et les paramètres λ et η. Pour chaque couple de paramètres (λ, η) fixé la meilleure carte, celle minimisant simultanément l'erreur de quantification vectorielle et l'erreur topologique a été retenue pour chaque table (cf. TAB 1). D'autres indices internes de performances tels que la mesure de distorsion auraient pu être utilisés.

	Structure en blocs		Structure de la carte			
Données	$\sharp blocs$	$\sharp VB$	Niter	Dim	$T_i \times T_f$	$(\lambda,\ \eta)$
IS	9-10		150	9×9	2×0.82	(3,31)
CT	7-4-10		150	10×10	2×0.1	(7,11)
DMU	76-216-64-240-47-6		150	10×7	3×0.2	(10, 20)
D1	25-25-25-25	9-18-10-7	150	10×9	2×0.1	(2, 3)
D2	5-5-5-5	2-2-4-4	150	10×9	3×0.2	(1, 5)
D3	5-5-5-5-5	2-2-4-4-5	150	10×10	3×0.2	(1, 5)

TAB. 1: Caractéristiques des données et paramètres des cartes retenus pour les tables IS, CT, DMU, D1, D2 et D3 ; il s'agit des cartes minimisant simultanément l'erreur topologique et de quantification vectorielle. Les quantités $\sharp blocs$ et $\sharp VB$ correspondent respectivement à la dimension de chaque bloc et au nombre de variables de bruit par bloc. Les quantités Niter, Dim, $T_i \times T_f$ et $(\lambda,\ \eta)$ correspondent respectivement au nombre d'itérations, aux dimensions de la carte, à la taille du voisinage et aux paramètres λ et η d'ajustement des poids α et β, les meilleures

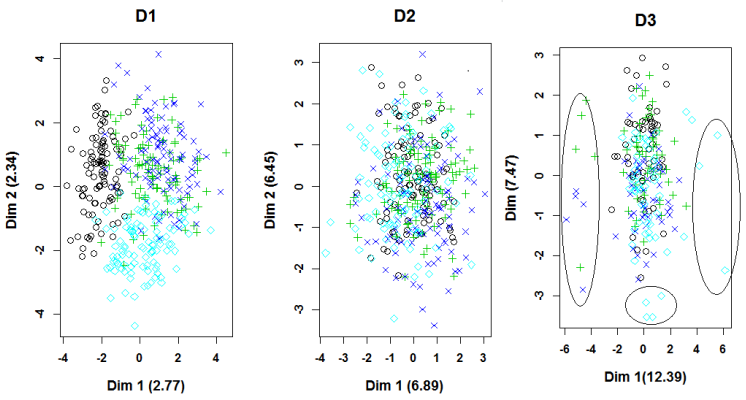


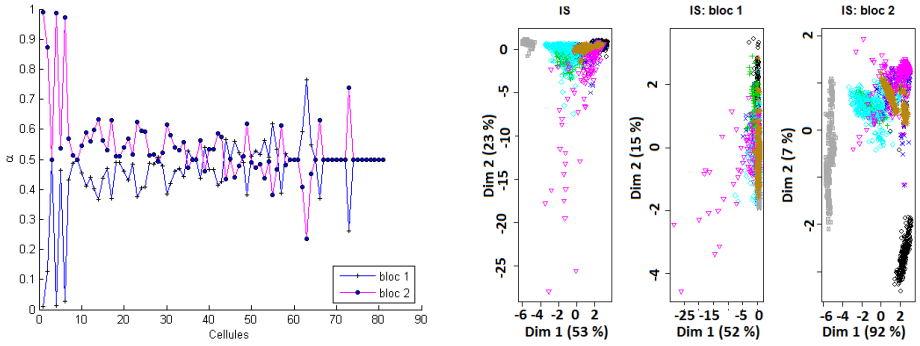
FIG. 1: Projection des classes des table D1, D2 et D3 dans le premier plan factoriel d'une ACP ; les observations atypiques sont entourées.

4.2 Résultats

Données réelles

Dans la suite, nous illustrons graphiquement la méthode uniquement sur la base IS, les résultats sont similaires pour les autres jeux de données.
La figure 2a donne une représentation graphique des poids α_{ck} définis sur les blocs par rapport aux cellules de la carte.
Pour les blocs, on observe que pour une majorité des cellules (plus de 67 %) les poids associés au bloc 2 sont nettement supérieurs à ceux du bloc 1. En d'autres termes, le bloc 2 apparaît donc plus pertinent pour déterminer la structure des classes de la carte. Une étude descriptive préalable de la table IS à travers une analyse en composantes principales permet, a priori, de

supposer ce résultat. En effet, la visualisation de la répartition des individus dans les classes sur le premier plan factoriel d'une ACP sur toutes les variables (figure 2b, IS) ou uniquement sur les variables de chaque bloc (figure 2b, IS : bloc 1) et (figure 2b IS : bloc 2), permet de voir que le bloc 2 permet de mieux distinguer les classes.

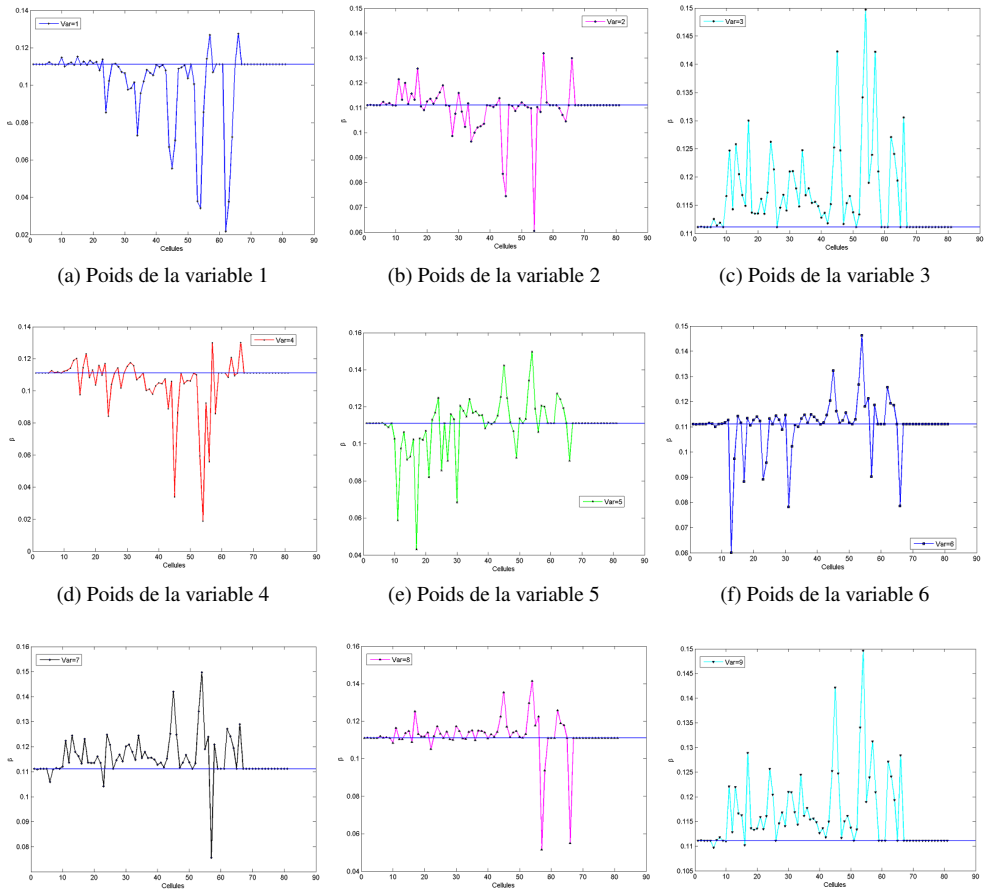


(a) Les poids α des blocs sur les cellules de la carte IS

(b) Projection des observations dans le premier plan factoriel défini par une analyse en composante principale sur la table IS

FIG. 2: Évaluation de la pertinence des blocs dans les cellules de la carte IS

Pour les variables, on formule, l'hypothèse qu'un poids β_{ckj} sur une variable v_j peut être considéré comme important s'il est supérieur au poids moyen $\frac{1}{p_k}$ de son bloc d'appartenance. Les figures 3 à 4 représentent les poids des variables en fonction des cellules de la carte topologique fournie par 2S-SOM. On observe, dans le bloc 1, au seuil $\frac{1}{p_1} = 0.11$, que les variables 3, 7, 8 et 9 sont influentes dans la plupart des cellules de la carte. Les variables 1, 2, 4 sont très peu influentes dans la plupart des cellules 30 à 50 et les variables 5 et 6 définissent les cellules comprises entre 50 et 60. Dans le bloc 2, les variables 1, 2, 3, 4 et 8 sont fortement influentes dans la majorité des cellules comprises entre 40 et 55. A l'opposée, les variables 7 et 10 ne sont pas pertinentes pour ces cellules. Plus précisément, le sous-espace associé à la cellule 40 de la carte IS par exemple est constitué des variables 1, 2, 3, 4, 8 et 9 du bloc 2 uniquement relativement au seuil 0.10 fixé (Cf. les figures 4a, 4b, 4c, 4d, 4e, 4i et 4j).
Nous illustrons maintenant les propriétés globales de 2S-SOM relativement aux paramètres λ et η. Les figures 5a et 5b représentent l'évolution de l'erreur de quantification vectorielle en fonction du couple λ et η. Lorsque les valeurs des paramètres λ et η augmentent, l'algorithme se stabilise au sens de l'erreur de quantification vectorielle ; leur influence sur cette dernière devient négligeable, 2S-SOM est alors semblable à SOM. Il apparaît deux valeurs particulièrement faibles assimilables à des "*outliers*" . Elle correspondent à un paramétrage d'apprentissage ($\lambda = 2$, $\eta = 16$,$\lambda = 11$, $\eta = 11$) qui dégrade la qualité topologique de la carte fournie par 2S-SOM. En effet la figure 9 présentée en annexe montre une erreur topologique associée non optimale (elle ne correspond pas à une valeur minimale) et des cellules formant la classe 1 qui ne sont pas toutes voisines sur la carte.
Les figures 5c et 5d représentent l'évolution des moyennes des poids des cellules de la carte associée à chaque paramètre λ lorsque η est fixé. (Nous l'illustrons avec λ =3 (figure 5e) et η



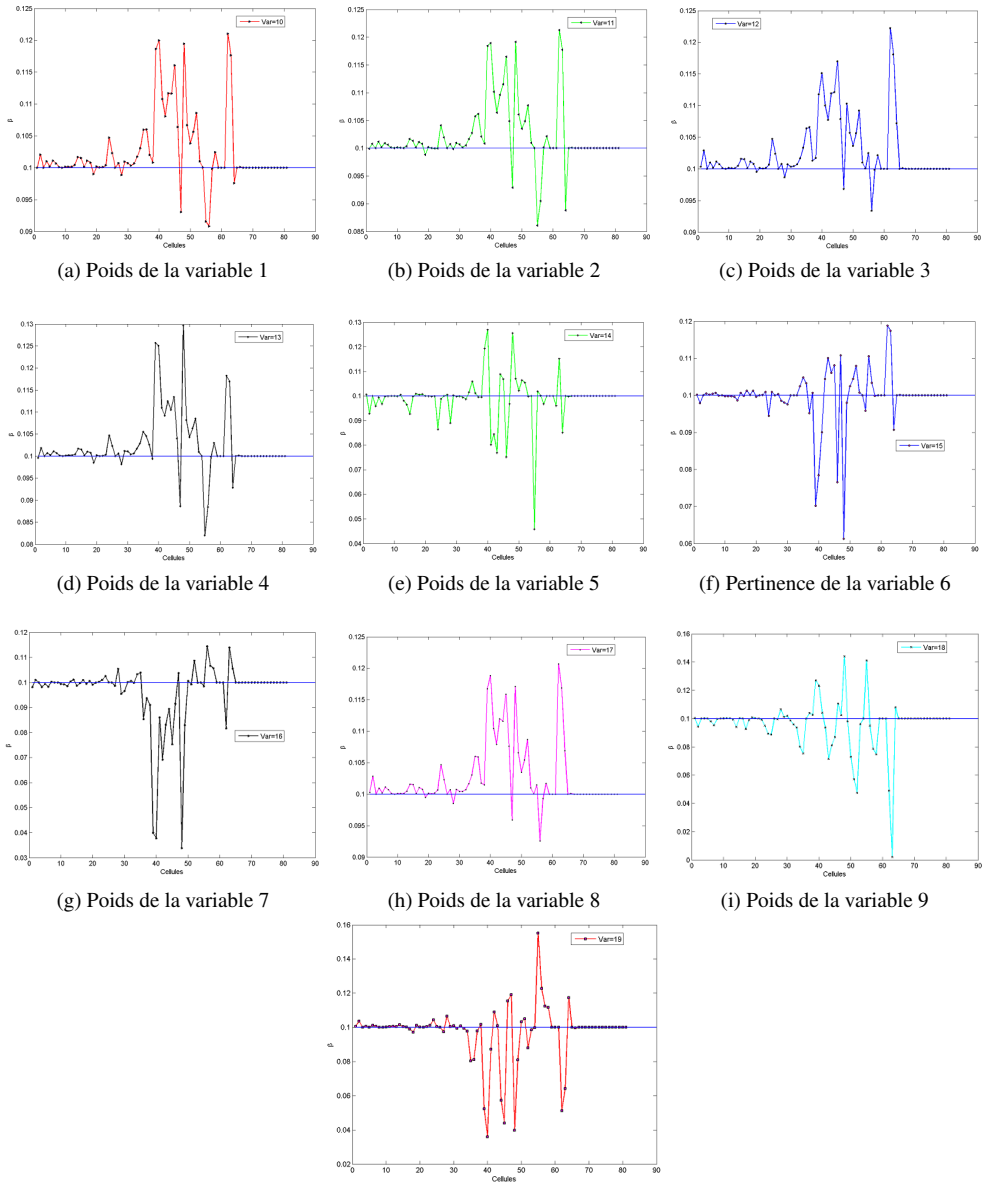
(a) Poids de la variable 1 (b) Poids de la variable 2 (c) Poids de la variable 3

(d) Poids de la variable 4 (e) Poids de la variable 5 (f) Poids de la variable 6

(g) Poids de la variable 7 (h) Poids de la variable 8 (i) Poids de la variable 9

FIG. 3: Représentation des poids β_{ckj} associés aux variables du bloc 1 par rapport au 81 cellules de la carte IS ; la ligne horizontale définit le seuil $\frac{1}{p_1} = 0.11$

= 3 (figure 5d)). On observe que lorsque $\lambda \to \infty$, les poids définis par l'algorithme 2S-SOM donnent les mêmes poids moyens aux blocs par cellule (cf. figure 5c). Tout se passe comme si on équilibrait l'influence des blocs, on ne prend en compte que les variables. Les poids moyens des variables définis par 2S-SOM pour l'ensemble des cellules de la carte permettent de regrouper les variables des blocs en groupes. Ainsi, dans le bloc 1 par exemple, les variables 3, 7 et 9 sont les plus pertinentes ($\beta_{ckj} > 0.11$) pour les cellules alors que les variables 4 et 5 sont moyennement pertinentes dans les cellules de la carte ($\beta_{ckj} \approx 0.11$) et les variables 1, 2, 6 et 8 aux poids inférieurs au seuil apportent de très faibles contributions dans les cellules de la carte. Il en est de même pour le bloc 2. (cf. figure 5e). Considérant chaque cellule de la carte lorsque $\lambda \to \infty$ pour η fixé, 2S-SOM permet de faire du subspace clustering en sélectionnant les variables ayant les plus forts poids pour la cellule.



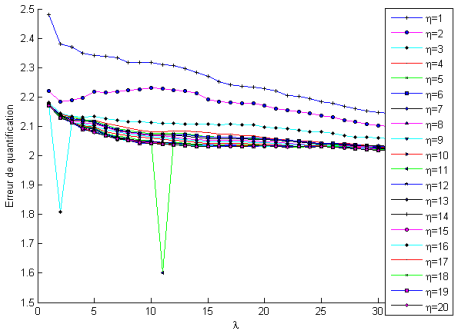
(a) Poids de la variable 1 (b) Poids de la variable 2 (c) Poids de la variable 3

(d) Poids de la variable 4 (e) Poids de la variable 5 (f) Pertinence de la variable 6

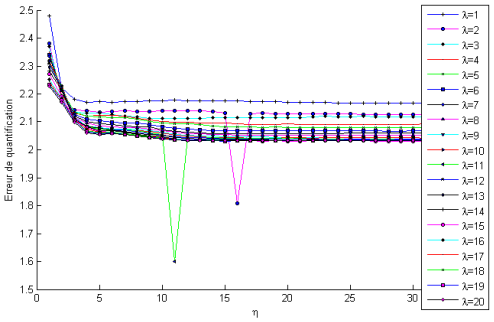
(g) Poids de la variable 7 (h) Poids de la variable 8 (i) Poids de la variable 9

(j) Poids de la variable 10

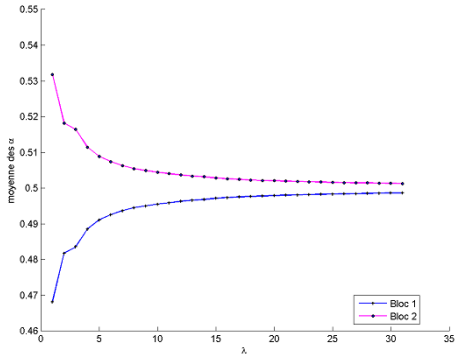
FIG. 4: Représentation des poids β_{ckj} associés aux variables du bloc 2 par rapport au 81 cellules de la carte IS ; la ligne horizontale définit le seuil $\frac{1}{p_2} = 0.10$



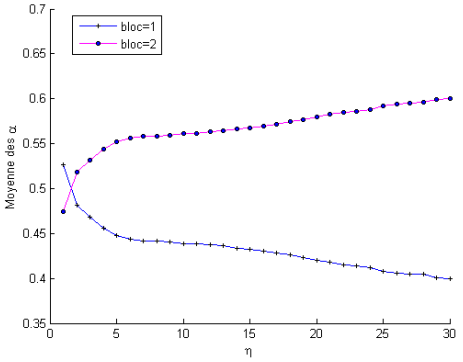
(a) L'évolution de l'erreur de quantification vectorielle par rapport au couple (λ, η), en abscisse les λ



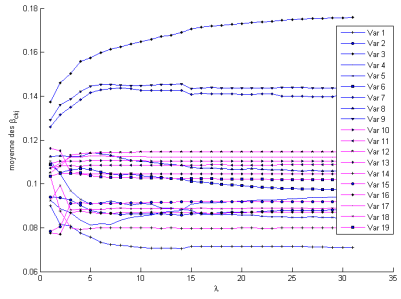
(b) L'évolution de l'erreur de quantification vectorielle par rapport au couple (λ, η), en abscisse les η



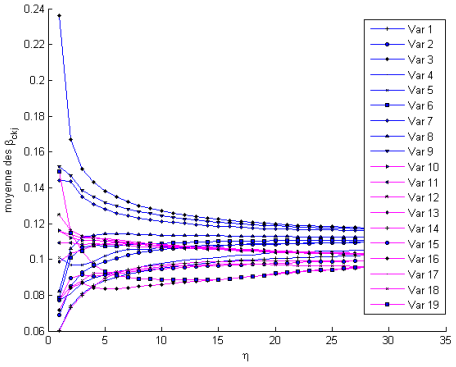
(c) L'évolution de la moyenne des poids des blocs sur les cellules pour les cartes obtenues avec λ variant et $\eta = 3$



(d) L'évolution de la moyenne des poids des blocs sur les cellules pour les cartes obtenues avec $\lambda = 3$ et η variant



(e) L'évolution de la moyenne des poids des variables sur les cellules pour les cartes obtenues avec λ variant et $\eta = 3$



(f) L'évolution de la moyenne des poids des variables sur les cellules pour les cartes obtenues avec $\lambda = 3$ et η variant

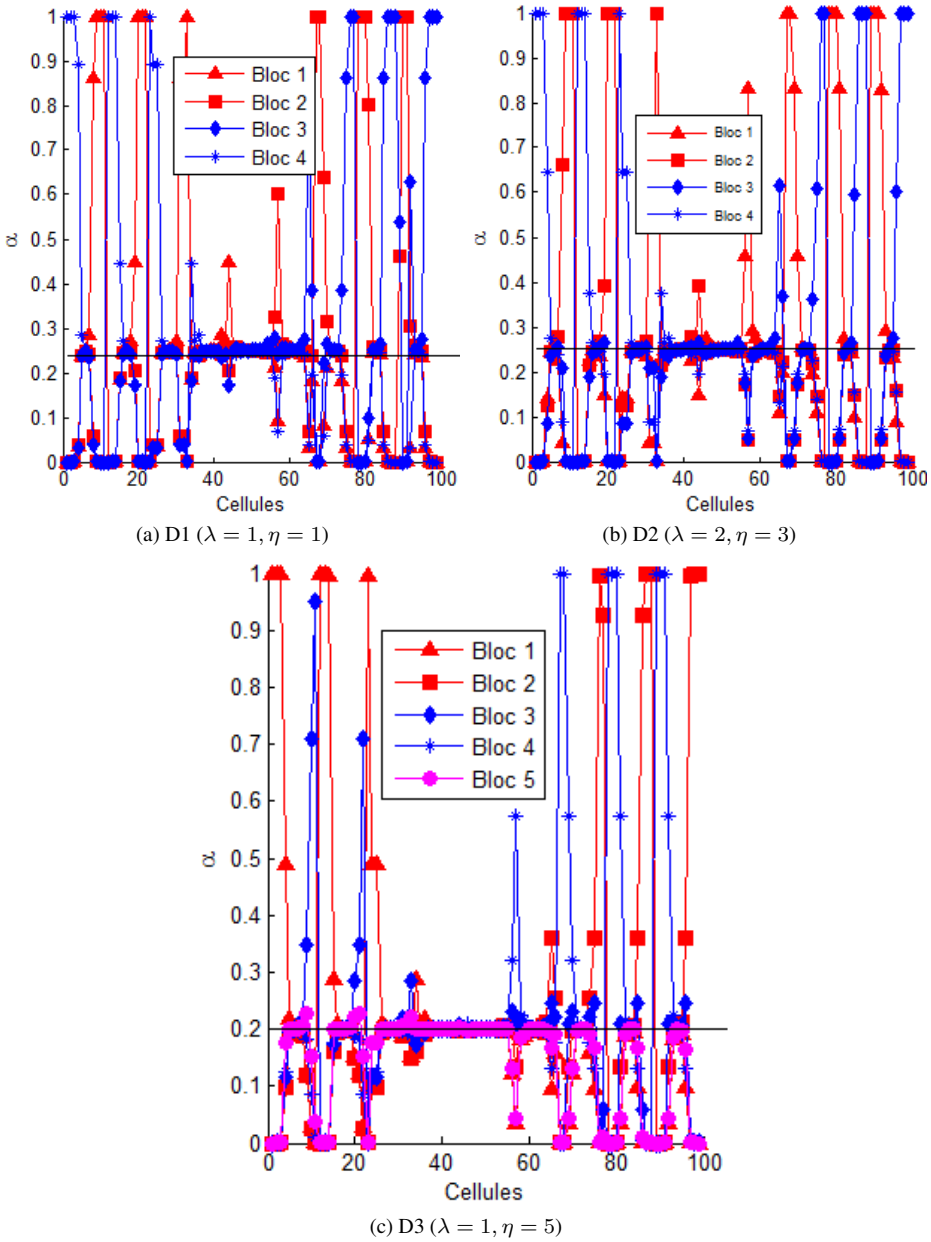
FIG. 5: Les propriétés de 2S-SOM par rapport aux paramètres λ et η

Lorsque λ fixé et $\eta \to \infty$, on observe des situations inverses : l'algorithme 2S-SOM fournit des poids quasi identiques pour toutes les variables et privilègie donc les blocs. On peut alors déterminer les blocs importants pour chaque cellule de la carte, on fait ainsi de la sélection de blocs plutôt que des variables séparément.

Données simulées

Les figures 6 représentent respectivement les poids α_{ck} attribués par 2S-SOM aux blocs dans les cellules des cartes topologiques associées aux tables D1, D2 et D3. Elles illustrent qu'à travers les poids forts chaque bloc se spécialise dans la définition d'un certain nombre de cellules. Sur la figure 6a par exemple, les blocs 1 et 4 caractérisent mieux les observations appartenant à la première moitié des cellules de la carte. Les blocs 2 et 3 définissent les observations appartenant à la deuxième moitié des cellules de la carte. Le constat est identique pour la table D2 avec les blocs 1 et 3 qui caractérisent les premières cellules et les blocs 2 et 4 les dernières cellules. Pour la table D3, on observe que les blocs 1, 2 contenant le moins de variables de bruit caractérisent un plus grand nombre de cellules parmi les premières et les dernières mais de façon exclusive. Les blocs 3 et 4 présentant plus de bruit donnent néanmoins des poids importants à un ensemble de cellules. Le bloc 5 porteur d'aucune information (il est composé uniquement de variables uniformes) n'est influent pour aucune cellule de la carte associée à la table D3 (Figure 6c). De manière générale, on observe que l'importance d'un bloc pour une cellule traduit la présence de variables fortement informatives pour la cellule par rapport aux autres blocs. En d'autres termes, la méthode proposée est robuste, par rapport à la présence de blocs contenant des variables de bruit.

Nous analysons dans la suite les poids des variables de la table D3 pour illustrer la capacité de l'algorithme à distinguer les variables de bruit et celles informatives. Dans les cellules ayant des poids forts pour le bloc 1 (figure 6c), les variables 1 et 5 (celles simulées avec une distribution uniforme) ont des poids faibles et ne sont donc pas prises en compte par 2S-SOM (figure 7a). Il en est de même pour les variables 1 et 3 du bloc 2 (figure 7b). Les blocs 3 et 4 mettent en évidence une seule variable pertinente (la seule à distribution non uniforme) pour les cellules influencées par ce bloc (figures 7c et 7d).
On constate aussi que dans le bloc 5 où toutes les variables ont une distribution uniforme, les poids sont quasi-identiques pour toutes les variables (Figure 7e).

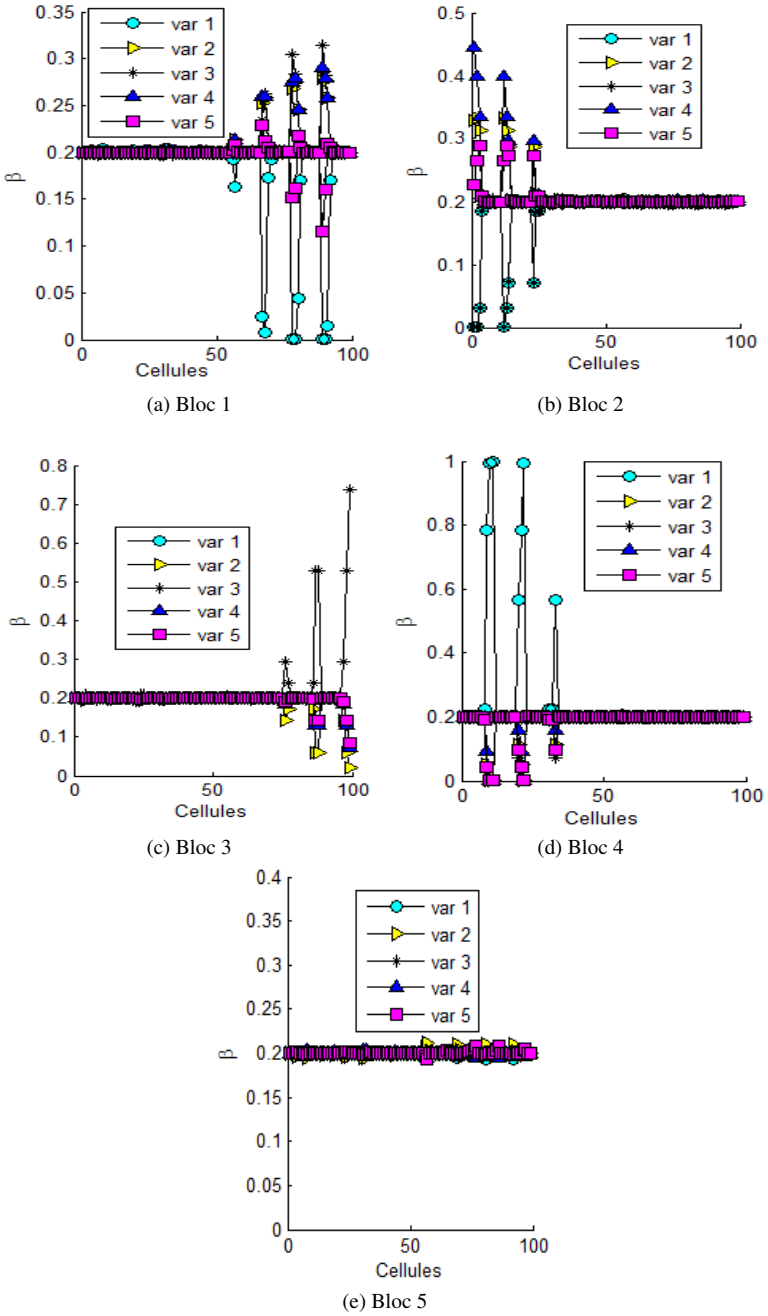


(a) D1 ($\lambda = 1, \eta = 1$)

(b) D2 ($\lambda = 2, \eta = 3$)

(c) D3 ($\lambda = 1, \eta = 5$)

FIG. 6: Les poids α_{ck} associés aux blocs par rapport aux cellules des cartes associées aux tables D1, D2 et D3



(a) Bloc 1

(b) Bloc 2

(c) Bloc 3

(d) Bloc 4

(e) Bloc 5

FIG. 7: Représentation des poids β des variables de la table D3 dans chaque bloc ($\lambda = 1$, $\eta = 5$)

4.3 Comparaison des performances

Nous utiliserons ici les indices externes de précision, de rappel, de F-mesure et le coefficient de pureté d'une partition sur des données labellisées. L'algorithme 2S-SOM a fournit des cartes de tailles relativement grandes (TAB 1). Une classification ascendante hiérarchique sous contraintes utilisant la stratégie d'agrégation de ward a été appliquée sur les référents des cartes finales pour obtenir des partitions en un nombre de classes identique au nombre de labels dans les jeux de données initiaux. Ainsi, les tables IS, DMU, CT10, D1, D2 et D3 ont été segmentées en 7, 10, 10, 4, 4 et 4 classes. Les indices externes de performances utilisés sont

Données	Cl 1	Cl 2	Cl 3	Cl 4
D1	100	100	99	101
D2	100	95	107	98
D3	237	146	10	7

TAB. 2: Répartition des observations dans les classes de la CAH

définis à partir d'un tableau de contingence pour deux partitions $\mathcal{C}$ et $\mathcal{C}'$. Aux observations z_i sont associés les paramètres suivants :

- N_{11} le nombre d'observations z_i, z_j qui sont dans les mêmes classes des partitions $\mathcal{C}$ et $\mathcal{C}'$.
- N_{00} le nombre d'observations z_i, z_j qui sont dans deux classes différentes de $\mathcal{C}$ et dans deux classes différentes de $\mathcal{C}'$.
- N_{10} le nombre d'observations z_i, z_j qui sont dans la même classe de la partition $\mathcal{C}$ et dans deux classes différentes de la partition $\mathcal{C}'$.
- N_{01} le nombre d'observations z_i, z_j qui sont dans deux classes différentes de la partition $\mathcal{C}$ et dans la même classe de la partition $\mathcal{C}'$.
- A_{ii} le nombre d'observations de la classe c_i de la partition $\mathcal{C}$ bien classées dans la classe c'_i de la partition $\mathcal{C}'$.

Précision	$Préc = \frac{N_{11}}{N_{11}+N_{01}}$
Rappel	$Rapp = \frac{N_{11}}{N_{11}+N_{10}}$
F-mesure	$Fm = \frac{2\times Prec\times Rapp}{Prec+Rapp}$
Pureté	$P = 100 * \frac{\sum_{k=1}^{3} A_{ii}}{N}$

TAB. 3: Critères d'évaluation des partitions

L'indice de précision et le coefficient de rappel sont des mesures asymétriques de similarité entre deux partitions dont l'une sert de référence. La précision indique la probabilité que deux objets soient regroupés dans une même classe de $\mathcal{C}'$ sachant qu'ils le sont dans la partition $\mathcal{C}$. Le coefficient de rappel indique la probabilité que deux objets soient regroupés dans la classe c sachant qu'ils le sont dans la classe c'. La F-mesure est définie comme la moyenne harmonique de l'indice de précision et du coefficient de rappel (TAB 3).

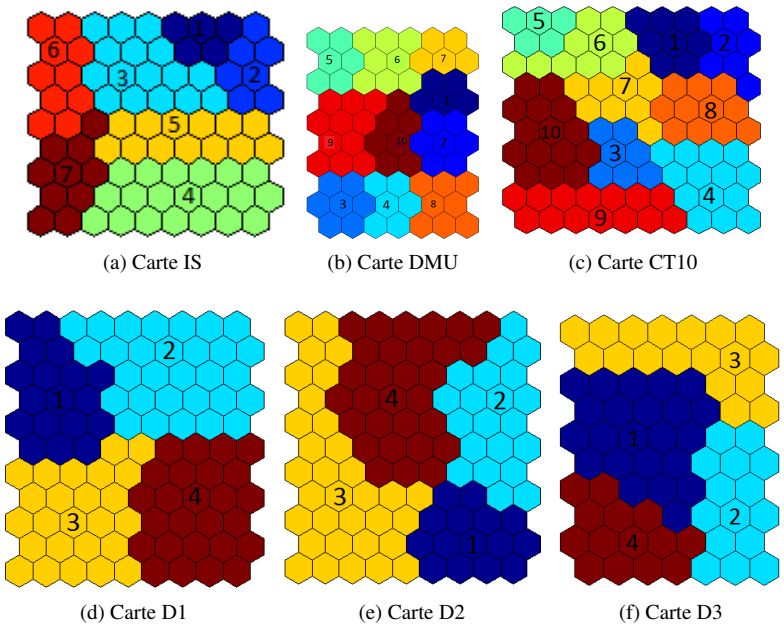
Les comparaisons des résultats (TAB 4) montrent de meilleures performances de 2S-SOM par rapport à FGKM sur tous les jeux de données et pour l'ensemble des indices de performance. Il en est de même pour 2S-SOM comparé à SOM et à EWKM à l'exception de la

mesure de rappel pour la base CT pour SOM et pour la base DMU pour EWKM. Comparé à la méthode des K-Means, 2S-SOM montre de meilleures performances sur les bases DMU et IS à l'exception du rappel pour la base IS. Pour la base CT, comparées aux classes des bases IS, DMU, D1, D2 et D3, les classes initiales sont déséquilibrées en termes de nombres d'observations (cf. TAB 5 en annexe), en particulier CT contient 4 classes de forts effectifs, que la stratégie d'agrégation de Ward utilisée ici ne permet pas de reconstituer convenablement. En effet, d'autres stratégies d'agrégation, notamment le lien minimum, permettent d'améliorer les performances en termes de précision (0.53), rappel (0.51), F mesure (0.50) et de pureté (0.50). Sur les données simulées D1, D2 et D3 la méthode 2S-SOM se révèle meilleure que l'ensemble des autres méthodes pour tous les indices. Les faibles performances des K-means et SOM sont peut être dues à l'incapacité de ces méthodes à ignorer les variables de bruit ou les blocs de bruit et à l'absence d'une structure globale de corrélation entre les variables. Comparée aux méthodes de type subspace clustering FGKM et EWKM la méthode proposée est meilleure en terme de performances pour tous les indices.

	Données	Indices	kM	EWKM	FGKM	SOM	2S-SOM
		Précision	0.38	0.66	0.60	0.63	**0.71**
		Rappel	**0.93**	0.70	0.63	0.67	0.74
	IS	F-mesure	0.50	0.64	0.59	0.59	**0.69**
		Pureté	0.41	0.59	0.63	0.61	**0.63**
		Precision	**0.50**	0.45	0.40	0.44	0.47
		Rappel	**0.53**	0.48	0.38	0.52	0.49
Réelles	CT	F-mesure	**0.48**	0.45	0.27	0.44	0.45
		Pureté	**0.47**	0.43	0.38	0.45	0.45
		Precision	0.59	**0.81**	0.60	0.75	0.80
		Rappel	0.61	**0.84**	0.80	0.78	0.82
	DMU	F-mesure	0.59	**0.80**	0.62	0.74	**0.80**
		Pureté	0.61	**0.77**	0.40	0.72	**0.77**
		Precision	0.37	0.98	0.90	0.31	**0.99**
		Rappel	0.35	0.65	0.60	0.28	**0.65**
	D1	F-mesure	0.36	0.77	0.77	0.29	**0.78**
		Pureté	0.47	0.72	0.72	0.38	**0.74**
		Précision	0.46	0.37	0.70	0.28	**0.85**
		Rappel	0.45	0.34	0.54	0.26	**0.55**
Simulées	D2	F-mesure	0.45	0.36	0.60	0.27	**0.66**
		Pureté	0.58	0.45	0.61	0.33	**0.66**
		Précision	0.33	0.35	0.75	0.35	**0.90**
		Rappel	0.28	0.30	0.48	0.27	**0.48**
	D3	F-mesure	0.31	0.36	0.61	0.29	**0.62**
		Pureté	0.37	0.47	0.49	0.35	**0.51**

TAB. 4: Performances des classifications de 2S-SOM les données réelles et sur les bases D1, D2 et D3

Au niveau de la visualisation, on observe une bonne conservation de la topologie des observations sur les cartes fournies par 2S-SOM sur chaque table. En effet, la méthode 2S SOM hérite des propriétés de SOM, en particulier les observations aberrantes sont isolées dans les classes 3 et 4 de la table D3 (Cf. figure 8).



(a) Carte IS (b) Carte DMU (c) Carte CT10

(d) Carte D1 (e) Carte D2 (f) Carte D3

FIG. 8: Visualisation des classes des cartes fournies par une CAH

5 Discussion

La méthode 2S-SOM proposée dans cet article permet de faire une classification d'un ensemble d'individus décrits par un grand ensemble de variables structurées en blocs et pouvant présenter des données aberrantes ou manquantes. Elle hérite des propriétés de visualisation de SOM et permet à travers les systèmes de poids associés de trouver des classes pertinentes ainsi que leurs sous espaces caractéristiques optimaux ; ce qui en facilite l'interprétation. Dans la littérature, plusieurs auteurs proposent des approches efficaces permettant de surmonter les difficultés liées aux données de grande dimension en classification. Kriegel et al. (2009) propose une revue détaillée de ces méthodes.
Une première approche consiste à réduire globalement la dimension des variables par application de méthodes d'analyse factorielle. L'approche, dite tandem, consiste alors à appliquer la classification sur un nombre optimal de composantes factorielles, dont le choix reste un problème délicat. Par ailleurs, plusieurs auteurs tels que (Hubert et Arabie, 1985; Vichi et Kiers, 2001) ont souligné les limitations de cette approche. En particulier, les composantes factorielles optimales pour maximiser l'inertie ne sont pas nécessairement dédiées à la découverte

des classes dans les données. Plusieurs méthodes permettent de lever ses limitations parmi lesquelles Reduced-K-Means (RKM) (De Soete et Carroll, 1994) and Factorial-K-Means (FKM) Vichi et Kiers (2001). Elles consistent à rechercher simultanément une classification des individus et une réduction des variables en composantes factorielles optimales pour la classification. Timmerman et al. (2010) effectue une étude comparative de ces méthodes alternatives à l'approche tandem. Cependant, les composantes factorielles sont des combinaisons linéaires de toutes les variables, ce qui en grande dimension ne permet pas toujours de faciliter l'interprétation en particulier lorsque les variables initiales ne sont pas fortement corrélées.
On peut aussi citer la méthode Local Dimensionality Reduction (LDR) (Chakrabarti et Mehrotra, 2000). L'idée consiste à rechercher des regroupements d'observations à l'aide d'une méthode de classification, puis à déterminer dans chaque groupe le sous espace associé par application d'une méthode factorielle sur les observations de la classe. La limitation majeure de la méthode LDR réside éventuellement dans le faible nombre d'observations dans certaines classes pour réaliser l'ACP. De plus, l'interprétation des classes grâce aux composantes factorielles demeure un problème.
Les travaux plus anciens d'Hartigan (1972) et de Govaert (1984) sur la recherche simultanée de groupes d'individus et de groupes de variables les décrivant par permutation directe des lignes et des colonnes du tableau initial des données, sont une alternative à la réduction local des dimensions (LDR). Govaert (1984) propose plusieurs approches de classification croisée liées au type de données. Le principe général consiste à définir alternativement une suite de couples de classifications (π_1^n, π_2^n) simultanément sur les lignes (π_1^n) et sur les colonnes (π_2^n) de la matrice initiale. Pour une partition des individus fixée, l'auteur recherche la partition des variables adaptée à cette partition des individus et vice-versa. La meilleure partition est alors choisie suivant le principe des nuées dynamiques en optimisant la distance entre le tableau de départ (partition triviale) et le tableau des partitions (π_1^n, π_2^n).
D'autres familles de méthodes de Bi-partitionnement issues de la théorie de l'information s'appuient d'une part sur la notion d'information mutuelle entre deux variables, d'autre part sur la mesure de divergence entre distribution de probabilités de Kullback-Leibler. Ces méthodes considèrent les deux partitions cherchées comme des variables aléatoires à valeurs discrètes et définissent la bipartition comme un problème de maximisation de l'association entre ces deux variables. Dhillon et al. (2003) utilisent la mesure de divergence entre distribution de probabilités de Kullback et Leibler et proposent de fixer a priori le nombre de classes de chacune des deux partitions puis optimisent localement une fonction objectif en estimant itérativement une partition en fonction de l'autre jusqu'à la convergence. A contrario, Robardet (2002) ne fixe pas a priori le nombre de classes des deux partitions et utilisent un algorithme d'optimisation locale stochastique qui procède également par ajustement itératif d'une partition en fonction de l'autre.
Le problème plus général du subspace clustering consistant à rechercher des classes dans différents sous-espaces de l'espace initial a été présenté par Agrawal et al. (1998); Parsons et al. (2004). Les auteurs proposent, des approches recherchant les sous-espaces à forte densité d'observations de l'espace initial. Ainsi, chaque dimension de l'espace est divisée en p intervalles de même longueur. Puis, un intervalle est alors considéré comme dense si le nombre d'observations qu'il contient est supérieur à un seuil fixé. Les sous-espaces denses définis sur k dimensions sont déterminés par le produit croisé des intervalles de chaque dimension et les groupes d'observations ne forment pas nécessairement une partition des données. Cependant,

ce type d'approche est limitée par la non-prise en compte des corrélations entre les variables. Pour atténuer la perte d'information rencontrée dans les méthodes de sélections des variables et de réduction globale des dimensions, Desarbo et al. (1984) proposent SYNCLUS, une méthode adaptée plutôt aux données de faible dimension, un algorithme de classification en deux étapes basé sur la méthode des K-Means avec des pondérations sur les variables. Elle consiste à déterminer k classes parmi les données initiales puis à estimer un système de pondération sur les variables en minimisant une certaine fonction quadratique moyenne associée aux k classes. Cette notion de pondération est étendue dans le cas de données de grande dimension : Domeniconi et al. (2004, 2007) propose dans Locally Adaptative Clustering (LAC), de définir les sous-espaces de variables associés aux classes à travers des pondérations locales sur les variables ; dans le même ordre d'idée, Huang et Ng (2005) dans W-K-Means puis Jing et al. (2007) dans Entropy weighting K-Means (EWKM) proposent dans de définir un système de pondération par modification de la fonction de coût associée à l'algorithme des K-Means en y introduisant des termes d'entropie. Les poids obtenus sont alors inversement proportionnels à la variance intra-classe des groupes, facilitant ainsi leur interprétation.
Lorsque le grand nombre de variables est, de plus, initialement structuré en blocs, le problème de la classification peut être reformulé comme la recherche d'un consensus entre les partitions obtenues sur chaque bloc séparément, appelées partitions initiales. Dans ces approches, les auteurs recherchent une partition consensus comme solution d'un problème d'optimisation basé sur un critère explicite liant cette partition aux partitions initiales associées à chaque bloc. On peut citer, par exemple, Strehl et Ghosh (2002); Topchy et al. (2005) qui utilisent le critère de l'information mutuelle normalisée et Gordon et Vichi (1998) qui se basent sur le critère de Rand.
Chen et al. (2012) proposent FGKM, une extension des approches de Huang et Ng (2005); Jing et al. (2007) à la classification des données de grande dimension initialement structurées en blocs. L'idée consiste à faire du soft-subspace clustering par la détermination des poids sur les variables et sur les blocs les plus informatifs pour chaque classe. FGKM est cependant basé sur la méthode des K-Means dont les performances dépendent fortement de l'initialisation des centres de classe.
Plus récemment, des travaux basés sur les cartes de Kohonen (1998) ont été proposés pour évaluer la contribution relative des variables dans les cellules d'une carte topologique. Dans le cas d'un seul bloc, Guérif et Bennani (2007) étend la méthode W-K-Means de Huang et Ng (2005) aux cartes topologiques auto-organisées en modifiant la fonction objectif de SOM par ajout d'un paramètre de pondération sur les variables.
Dans le cas de plusieurs blocs, Niang et Ouattara (2013) proposent HMTM, une méthode de type hiérarchique à deux niveaux pour trouver un consensus de partition basé sur des variables mixtes. Elle consiste à appliquer la même méthode des cartes topologiques mixtes (MTM) (Lebbah et al., 2005) pour obtenir, au premier niveau, une classification par bloc, puis dans une seconde étape pour rechercher une classification consensus des partitions préalables. Cependant, le niveau 1 de HMTM hérite des problèmes liés aux données de grande dimension et le niveau 2 ne prend pas en compte les différences entre les partitions initiales. Ouattara et Niang (2012) proposent une version basée sur SOM de la méthode HMTM pour prendre en compte, au niveau 2, les contributions relatives des blocs dans les cellules de la carte topologique.

6 Conclusion

La méthode 2S-SOM proposée dans cet article permet de faire une classification d'un ensemble d'individus décrits par un grand ensemble de variables structurées en blocs et pouvant présenter des données aberrantes ou manquantes. Son application sur des données étiquetées fournit des partitions, globalement, plus en adéquation avec les partitions de référence que celles obtenues avec les méthodes SOM, K-Means, EWKM et FGKM. Des travaux sont en cours sur des évaluations plus formelles des résultats et sur l'étude de leur robustesse. Par ailleurs, nous envisageons une application de 2S-SOM sur des données réelles de l'observatoire de la qualié de l'air intérieur.

Références

Agrawal, R., J. Gehrke, D. Gunopulos, et P. Raghavan (1998). Automatic subspace clustering of high dimensional data for data mining applications. pp. 94–105.

Chakrabarti, K. et S. Mehrotra (2000). Local dimensionality reduction : A new approach to indexing high dimensional spaces. In *Proceedings of the 26th International Conference on Very Large Data Bases*, pp. 89–100. Citeseer.

Chen, X., Y. Ye, et al. (2012). A feature group weighting method for subspace clustering of high-dimensional data. *Pattern Recogn 45(1)*, 434–446.

Cottrell, M., P. Letrémy, P. Rousset, S. Ibbou, et al. (2003). Cartes auto-organisées pour l'analyse exploratoire de données et la visualisation. *Journal de la société française de statistique 144*, 67–106.
eng

De Soete, G. et J. Carroll (1994). K-means clustering in a low-dimensional euclidean space.

Desarbo, W., J. Carroll, L. Clark, et P. Green (1984). Synthesized clustering: A method for amalgamating clustering bases with differential weighting variables. *Psychometrika 49*, 57–78.

Dhillon, I. S., S. Mallela, et D. S. Modha (2003). Information-theoretic co-clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 89–98. ACM.

Domeniconi, C., D. Gunopulos, S. Ma, B. Yan, M. Al-Razgan, et D. Papadopoulos (2007). Locally adaptive metrics for clustering high dimensional data. *Data Mining and Knowledge Discovery 14*, 63–97. 10.1007/s10618-006-0060-8.

Domeniconi, C., D. Papadopoulos, D. Gunopulos, et S. Ma (2004). Subspace clustering of high dimensional data. In *SIAM Int. Conf. on Data Mining*.

Gordon, A. (1996). A survey of constrained classification. *Computational Statistics & Data Analysis 21*(1), 17 – 29.

Gordon, A. et M. Vichi (1998). Partitions of partitions. *Journal of Classification 15*(2), 265–285.

Govaert, G. (1984). Classification simultanée de tableaux binaires. *In E. Diday, M. Jambu, L. Lebart, J. Pages, and R. Tomassone, editors, Data analysis and informatics III,North*

Holland 67(337)(1), 233–236.

Guérif, S. et Y. Bennani (2007). Sélection de variables en apprentissage numérique non supervisé. Conférence francophone sur l'Apprentissage automatique', Grenoble, France.

Hartigan (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association 67(337)*(1), 123–129.

Huang, J. Z. et M. K. Ng (2005). Automated variable weighting in k-means type clustering. *IEEE Transaction on pattern analysis and machine intelligence 27(5)*, 657–668.

Hubert, L. et P. Arabie (1985). Comparing partitions. *Journal of classification 2*(1), 193–218.

Jing, L., M. Ng, et J. Huang (2007). An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *Knowledge and Data Engineering, IEEE Transactions on 19*(8), 1026 –1041.

Kaski, S. (1997). Data exploration using self-organizing maps.

Kirchner, S., B. Andrée, C. Cochet, C. Dassonville, M. Derbez, Y. Leers, J. Lucas, C. Mandin, O. Ramalho, et J. R. M. Ouattara (2011). *Qualité d'air intérieur, qualité de vie. 10 ans de recherche pour mieux respirer*. paris: CSTB éditions.

Kohonen, T. (1998). The self-organizing map. *Neurocomputing 21(1-3)*.

Kohonen, T. (1999). Comparison of som point densities based on different criteria. *Neural Comput. 11*(8), 2081–2095.

Kriegel, H.-P., P. Kröger, et A. Zimek (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowl. Discov. Data 3*, 1:58.

Lebbah, M., A. Chazotte, F. Badran, et S. Thiria (2005). Mixed topological map. *ESANN 17*.

Niang, N. et M. Ouattara (2013). Hierarchical mixed topological maps. *Revue des Nouvelles Technologies de l'Information Advances in Theory and Applications of High Dimensional and Symbolic Data Analysis, RNTI-E-25*, 123–138.

Ouattara, M. et N. Niang (2012). Classification multi blocs pondérée basée sur les cartes topologiques auto-organisées (consom). In *XIX conférence de la Société Française de classification SFC : 2012*, Marseille.

Parsons, L., E. Haque, et H. Liu (2004). Subspace clustering for high dimensional data: a review. *SIGKDD Explor. Newsl. 6*(1), 90–105.

Robardet, C. (2002). Contribution à la classification non supervisée: proposition d'une methode de bi-partitionnement. *PhDThesis Université Claude Bernard - Lyon 1*, 311–331.

Strehl, A. et J. Ghosh (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research 3*, 583–617.

Timmerman, M. E., E. Ceulemans, H. A. Kiers, et M. Vichi (2010). Factorial and reduced k-means reconsidered. *Computational Statistics & Data Analysis 54*(7), 1858 – 1871.

Topchy, A., A. K. Jain, et W. Punch (2005). Clustering ensembles: Models of consensus and weak partitions. *Ieee Transaction on pattern analysis and machine intelligence*, 1866–1881.

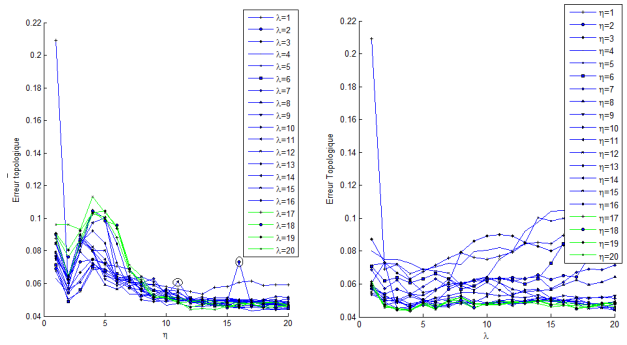
Vesanto, J., J. Himberg, E. Alhoniemi, et J. Parhankangas (2000). *SOM toolbox for Matlab 5*. Citeseer.

	Cl	1	2	3	4	5	6	7	8	9	10
IS	$\sharp Cl$	330	330	330	330	330	330	330			
DMU	$\sharp Cl$	200	200	200	200	200	200	200	200	200	200
CT	$\sharp Cl$	384	579	53	81	72	332	252	107	69	197

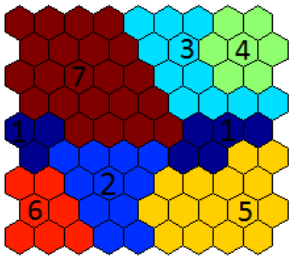
TABLE 5: Proportion d'observations par classe

Vichi, M. et H. A. Kiers (2001). Factorial k-means analysis for two-way data. *Computational statistics & data analysis 37*(1), 49–64.

Annexe



(a) Erreur topologique (b) Erreur topologique



(c) Les classe après une CAH sur les référents de la carte fournie par *outlier*

FIGURE 9: Représentation de l'erreur topologique pour la carte IS et la carte associée aux référents

Soft Subpace clustering basé sur SOM

Summary

We propose 2S-SOM, a soft subspace clustering method based on self organizing maps (SOM) for clustering individuals described by several variables structured in homogeneous blocks. The proposed algorithm optimizes the cost function of SOM modified by introducing adaptative weights on the blocks and on the variables. This double weighting allows the identification of significant blocks and significant variables in each block and thereby take into account the variable blocking.
2S-SOM allows to identify simultaneously clusters or groups of objects and their optimal subspace of clustering. The method is illustrated on real data from UCI repository of machine learning and on simulated data sets