Two step soft subspace SOM : une méthode de classification multi-bloc avec sélection de variables

François Kaly, Ndèye Niang, Mory Ouattara, Niang Awa, Sylvie Thiria, Beatrice Marticorena, Serge Janicot

Two step soft subspace SOM : une méthode de classification multi-bloc avec sélection de variables

F. Kaly[*,**,***] Ndèye Niang[****]
Mory Ouattara[****] Awa Niang[*] Sylvie Thiria[***] Beatrice Marticorena[**] Serge Janicot[***]

[*]LTI, UCAD, Dakar, Sénégal

[**] LISA, UMR CNRS 7583 ; IPSL ; UPEC ; UPD, Créteil, France

[***]LOCEAN, UMR 7159 CNRS-IRD-UPMC-MNHN; IPSL, Paris, France

[****]Statistique Appliquée, CNAM 292, rue Saint Martin, 75141 Paris Cedex 03, France

Résumé. Nous proposons une stratégie de classification de données multi-blocs basée sur l'utilisation de la méthode de soft subspace clustering 2S-SOM dans un processus hiérarchique à deux niveaux combiné à des tests statistiques. Une première application de la méthode 2S-SOM fournit un système de poids évaluant les contributions relatives des variables et des blocs aux groupes d'observations. Nous proposons une procédure de test statistique permettant de sélectionner les variables significativement pertinentes. 2S-SOM est à nouveau utilisée sur ces dernières pour déterminer la partition finale des observations. La méthode est évaluée sur des données simulées et réelles. En particulier, l'application sur des données météorologiques montre que la sélection des variables au niveau 1 facilite l'interprétation des classes obtenues.

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 termes de 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. De plus ces données peuvent présenter une structure en plusieurs blocs de variables caractérisant chacune une vue particulière sur les données recueillies selon une thématique spécifique, on parle de données multi-vues ou multi-blocs. Or, la majorité des mesures de distance perdent leur pouvoir discriminant au fur et à mesure que la dimension augmente ; les observations étant pratiquement toutes équidistantes les unes par rapport aux autres, Parsons et al. (2004). En outre, en l'absence d'une structure globale de corrélation entre les variables (ce qui est souvent le cas en grande dimension à cause de 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 à rechercher dans des sous-espaces de l'espace initial, on parle alors de méthode de subspace clustering. Le principe des méthodes de 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.; Kriegel et al. (2009). Des approches récentes basées sur l'introduction dans la méthode des K-moyennes d'une pondération des variables ou des blocs permettent de prendre en compte en plus de la grande dimension, la structure multi-blocs (Jing et al., 2007; Chen et Ye, 2012). Ce sont des méthodes de soft subspace clustering. Plus récemment, ces approches ont été étendues, à travers une nouvelle méthode 2S-SOM Ouattara et al. (2014), aux cartes auto-organisées ou self organizing maps (SOM) Kohonen (1998) permettant ainsi d'exploiter les propriétés de visualisation de SOM.
Nous proposons une approche hiérarchique de sélection de variables en classification fondée sur une double utilisation de 2S-SOM. Dans une première application, 2S-SOM fournit un système de poids à partir duquel on recherche des variables ou des blocs pertinents à l'aide d'un test d'hypothèse statistique. Ensuite, une deuxième utilisation de 2S-SOM sur les variables sélectionnées fournit la partition recherchée. La méthode proposée est présentée en section 2 après les notations et la description de 2S-SOM. La section 3 présente son illustration sur des données réelles. La section 4 est consacrée à la conclusion.

2 Soft-Subspace clustering basé sur SOM : 2S-SOM

Nous disposons de N observations z_i décrites par p variables divisées en B blocs. On recherche une partition des observations en K classes.
Les notations suivantes seront utilisées :

- $\mathcal{V} = \{z^j, j = 1, \ldots, p\}$ l'ensemble des variables divisé en B blocs de p_b variables tels que $p_1 + \ldots + p_b + \ldots + p_B = p$.
- α est une matrice $K \times B$ où K désigne le nombre de classes c dans Z, α_{cb} est le poids du bloc b dans la classe c de Z.
- $\beta = [\beta_1, \ldots, \beta_B]$ est une matrice $K \times p$ où β_b est une matrice de dimension $K \times p_b$ définissant les poids $\beta_{cbj} (j = 1, \ldots, p_b)$ sur les variables du bloc b pour chaque c de Z.

2.1 2S-SOM

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 K cellules c structurées par des graphes non-orientés permettant de définir a priori une distance entre les 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 initial des

cartes topologiques consiste à minimiser de manière itérative en deux phases la fonction de coût (Kohonen, 1998) :

$$\mathcal{J}^T_{SOM}(\mathcal{Z},\mathcal{W}) = \sum_{i=1}^{N}\sum_{c\in C}\mathcal{K}^T(\sigma(\mathcal{X}(z_i),c))\sum_{j=1}^{p}(z_{ij}-\omega_{cj})^2 \quad (1)$$

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.
Dans le cas des données en bloc, l'approche de type subspace clustering 2S-SOM repose sur une modification de la fonction de coût de SOM en introduisant un double système de poids $\alpha_{cb}(b = 1, \ldots, B)$ et $\beta_{cbj}(j = 1, \ldots, p_k)$ définis respectivement sur les blocs et sur les variables pour chaque cellule $c \in C$. La classification et les poids relatifs à la pertinence des blocs et des variables sont donc obtenus par minimisation de la fonction objectif J^T_{2SSOM} définie par la relation suivante :

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

avec $d_{\beta_{cb}} = \sum_{j=1}^{p_b}\beta_{cbj}(z_{ibj} - \omega_{ckj})^2$ et sous les contraintes :

$$\begin{cases} \sum_{j=1}^{p_b}\beta_{cbj} = 1,\ \beta_{cbj}\in[0,1], \forall\, c\in\mathcal{C}, \forall b \\ \qquad \sum_{b=1}^{B}\alpha_{cb} = 1,\ \alpha_{cb}\in[0,1],\ \forall\, c\in\mathcal{C} \end{cases} \quad (3)$$

$I_c = \lambda\sum_{b=1}^{B}\alpha_{cb}log(\alpha_{cb})$ et $J_{cb} = \eta\sum_{j=1}^{p_b}\beta_{cbj}log(\beta_{cbj})$ représentent des termes entropies négatives pondérées et associées aux vecteurs poids relatifs aux blocs et aux variables pour une cellule c. La minimisation de la fonction de coût J^T_{2S-SOM} s'effectue de façon alternée en quatre étapes dont les deux premières phases d'affectation des observations aux classes et d'actualisation des vecteurs référents sont identiques à celles de la méthode SOM. 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) :

$$\mathcal{X}(z_i) = \underset{c\in\mathcal{C}}{argmin}\left(\sum_{r\in\mathcal{C}}\left(\sum_{b=1}^{B}\alpha_{cb}\mathcal{K}^T(\sigma(r,c))d_{\beta_{cb}}\right)\right) \quad (4)$$

— étape 2 : Actualisation des centres de classe à l'aide de : (5)

$$\omega_c^T = \frac{\sum_{i=1}^{N} \mathcal{K}^T(\sigma(\mathcal{X}(z_i), c))z_i}{\sum_{i=1}^{N} \mathcal{K}^T(\sigma(\mathcal{X}(z_i), c))} \tag{5}$$

A l'aide du lagrangien de la fonction J_{2S-SOM}^T relativement aux quantités α et β on détermine les poids α et β associés respectivement aux blocs et aux variables. Ainsi, on obtient :

— Etape 3 : les paramètres $\mathcal{X}$, ω et β connus et fixés à leurs valeurs courantes alors on a :

$$\alpha_{cb} = \frac{exp(\frac{-\Psi_{cb}}{\lambda})}{\sum_{b=1}^{B} exp(\frac{-\Psi_{cb}}{\lambda})} \tag{6}$$

avec

$$\Psi_{cb} = \sum_{z_i \in r, r \neq c} \mathcal{K}^T(r, c) d_{\beta_{cb}} + \mathcal{K}^T(c, c) \sum_{z_i \in c} d_{\beta_{cb}} \tag{7}$$

— Etape 4 : de manière identique, si les paramètres $\mathcal{X}$, ω et α sont connus et fixés à leurs valeurs courantes alors , on a :

$$\beta_{cbj} = \frac{\exp(\frac{-\Phi_{cbj}}{\eta})}{\sum_{j=1}^{p_b} \exp(\frac{-\Phi_{cbj}}{\eta})} \tag{8}$$

avec

$$\Phi_{cbj} = \sum_{z_i \in r, r \neq c} \alpha_{cb} \mathcal{K}^T(r, c)(z_{ibj} - \omega_{cbj})^2 + \mathcal{K}^T(c, c) \sum_{z_i \in c} \alpha_{cb}(z_{ibj} - \omega_{cbj})^2 \tag{9}$$

Le poids d'une variable ou d'un bloc sera donc d'autant plus important qu'il minimise simultanément la somme des écarts entre les référents w_c et les observations appartenant à la cellule c et aux cellules r du voisinage T de la cellule c. Les coefficients de pondération α_{cb} et β_{cbj} définis par 2S-SOM indiquent respectivement l'importance relative des blocs et des variables dans les classes. Ainsi, plus le poids d'un bloc b ou d'une variable 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.

2.2 Sélection des variables pertinentes

Au niveau des cellules, la pertinence d'une variable ou d'un bloc est fournie directement par son poids défini par 2S-SOM.
Les poids α_{cb} et β_{cbj} sont définis pour un bloc b et pour une cellule c sous les contraintes

$\sum_j^{p_b} \beta_{cbj} = 1$ et $\sum_b^B \alpha_{cb} = 1$. La contribution moyenne d'une variable à une cellule de la carte est donc $\frac{1}{p_b}$ que l'on peut utiliser comme un seuil de sélection des variables pertinentes. Une variable telle que $\beta_{cbj} < \frac{1}{p_b}\ \forall\ c \in \mathcal{C}$ sera non- pertinente donc éliminée.

Dans certains cas, la taille de la carte peut conduire à un grand nombre de cellules. Il est alors 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 $K^{'}$ de classes contenant $n_{k'}$ cellules par classe k' (Gordon, 1996).

Ce regroupement des cellules engendre la nécessité de définir le poids et la pertinence d'une variable dans les classes. Nous proposons de prendre la moyenne des poids des cellules constituant la classe, soit $\gamma_{k'bj} = \frac{1}{n_{k'}} \sum_{l=1}^{n_{k'}} \beta_{c_l^{k'}bj}$ pour la variable j du bloc b et la classe k' composée des cellules $c_l^{k'}$.

De même, on définit $\delta_{k'b} = \frac{1}{n_{k'}} \sum_{l=1}^{n_{k'}} \alpha_{c_l^{k'}b}$ la moyenne des poids du bloc b sur les cellules de la classe k'.

Pour évaluer la pertinence d'une variable dans une classe de la CAH, nous proposons d'utiliser un test statistique basé sur le principe de la valeur-test proposée par Lebart et al. (1997). On désigne par $\gamma_{bj} = \frac{1}{N_{cell}} \sum_{c \in \mathcal{C}} \beta_{cbj}$, s_{bj}^2, $s_{k'bj}^2 = \frac{N_{cell} - n_{k'}}{N_{cell} - 1} \frac{s_{bj}^2}{n_{k'}}$ respectivement la moyenne, la variance des poids de la variable j du bloc b pour l'ensemble des cellules et la variance des poids de la variable j du bloc b dans la classe k' de la CAH. La valeur-test, pour les poids d'une variable j dans une classe k', se définit alors par :

$$t_{k'bj} = \frac{\gamma_{k'bj} - \gamma_{bj}}{s_{k'bj}}$$

La valeur test $t_{k'bj}$ peut se lire comme la statistique d'un test de comparaison de moyennes où, sous l'hypothèse nulle de tirage au hasard de $n_{k'}$ cellules parmi N_{cell}, elle suivrait de manière asymptotique la loi normale centrée réduite. Pour les niveaux de risque usuels (5%), on considérera donc que la différence entre le poids moyen dans la classe et le poids moyen sur la carte est significative lorsque la valeur absolue de la valeur test est supérieure à 2.

Ainsi, au niveau de la valeur test, la pertinence d'une variable ou d'un bloc dans une classe de la CAH est relative à la distribution des poids de 2S-SOM pour l'ensemble des cellules de la carte.

Pour une classe la contribution moyenne des variables est :

$$\frac{1}{p_b} \sum_{j=1}^{p_b} \gamma_{k'bj} = \frac{1}{p_b} \sum_{j=1}^{p_b} \frac{1}{n_{k'}} \sum_{l=1}^{n_{k'}} \beta_{c_l^{k'}bj} = \frac{1}{p_b}$$

Par ailleurs, compte tenu des propriétés de 2S-SOM (Ouattara et al., 2014), les variables de bruit qui ont en général des poids $\gamma_{bj} < \frac{1}{p_b}$ peuvent également être sélectionnées par la valeur test. Ainsi, parmi les variables sélectionnées par la valeur test, seront éliminées celles dont le poids moyen $\gamma_{bj} < \frac{1}{p_b}$.

Il est alors possible de sélectionner l'ensemble des variables pertinentes pour les blocs à travers une première application de la méthode 2S-SOM.

L'approche hiérarchique que nous proposons ici, consiste ensuite, à appliquer de nouveau

2S-SOM sur les variables sélectionnées pour obtenir la partition recherchée. La réduction du nombre de variables permet de simplifier d'une part l'interprétation des classes obtenues, d'autre part elle fournit des partitions de meilleure qualité en terme d'indices de pureté et de NMI comme cela est illustré ci-dessous sur des données labellisées. Les notions de pureté et de NMI sont detaillées dans l'annexe. sNous évaluons les performances de notre approche de sélection des variables sur un jeu de données réelles issus de l'UCI et sur un jeu de données simulées :

— Le jeu de données "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. Chaque observation possède une étiquette comprise entre 1 et 7.

— Les données simulées D contiennent 400 observations divisées en 4 classes de 100 observations décrites par 4 blocs de variables. Elle contiennent 4 blocs de 5 variables. Les blocs contiennent respectivement 2, 2, 4 et 4 variables de bruit.

Le tableau 1 présente la valeur moyenne des performances de 2S-SOM au niveau 1 (avant sélection), au niveau 2 (après sélection) grâce au test d'hypothèse et par rapport à la sélection brute dans laquelle une variable est jugée importante si son poids est supérieur à $1/p_b$.

Data	Index		SOM	2S-SOM	2S-SOM$_{VT}$	EWKM	FGKM
D	NMI	mu	0.11	0.82	0.81	0.43	0.32
		std	0.04	0.08	0.06	0.01	0.34
	Pureté	mu	0.35	0.89	0.90	0.32	0.56
		std	0.06	0.10	0.11	0.01	0.22
IS	NMI	mu	0.60	0.60	0.64	0.53	0.40
		std	0.02	0.08	0.06	0.07	0.14
	Pureté	mu	0.61	0.61	0.64	0.61	0.63
		std	0.03	0.06	0.05	0.05	0.05

TAB. 1 – *Comparaisons des performances des méthodes 2S-SOM avant (2S-SOM) et après la sélection des variables (2S-SOM$_{VT}$) avec les performances des méthodes basées sur la méthode des K-moyennes (EWKM, FGKM)*

	C_1	C_2	C_3	C_4
Bloc	(2,3)	3	4	1
Var	(2-3, 3)	0	1	3

TAB. 2 – *Variables pertinentes*

Le tableau 2 montre qu'on sélectionne effectivement les variables non-informatives pour la classification puisque la suppression des variables non-pertinentes ne dégrade pas les performances de classification sur les bases D et IS. Par ailleurs, les performances de la méthodes 2S-SOM$_{VT}$ restent meilleures que celles des méthodes EWKM, FGKM.

3 Application

3.1 Données

Les données utilisées, sont les champs météorologiques ré-analysées ERAInterim du centre européen de prévision météorologique (ECMWF). Les observations sont fournies toutes les 3 heures selon des grilles superposées en altitude correspondant aux 9 niveaux de pressions (1000 hpa, 925 hpa, 850 hpa, 700 hpa, 500 hpa, 400 hpa, 300 hpa, 250 hpa, 200hpa) et dont chaque maille carrée de 0.5° de coté décrit l'état de l'atmosphère. Ces données sont initialement constituées de 7 paramètres météorologiques que sont la température (T), l'humidité spécifique (Q), le géopotentiel (Z), le vent zonal (U) et le vent méridien (V), la vitesse verticale du vent (W) et la hauteur de la couche limite (BLH) qui servent à la classification des données. Chacun des 6 premiers blocs initiaux est constitué de 153 variables et 2549 observations correspondant à l'état de l'atmosphère en une heure donnée. Le bloc BLH est constitué de 17 variables. L'étude porte sur les données de la saison sèche (d'Octobre à Mai) des années allant de 2006 à 2010 de la station de Mbour (Sénégal). Ainsi, on a formé une base de données de 2549 observations sur 935 variables. Dans ce travail, les blocs sont structurés selon deux thématiques : l'une spécifique à la thermodynamique de l'atmosphère, il s'agit des paramètres T, U, V et Z composant un bloc de 612 variables et le deuxième bloc correspond aux mouvements de l'atmosphère (Q, W, BLH) et est composé de 323 variables. Une carte topologique a été réalisée avec 2S-SOM introduisant pour chaque bloc et pour chaque variable un système de poids adaptatifs. A partir de ces poids on identifie les variables et les blocs les plus pertinents dans la classification.

3.2 Résultats

L'algorithme a été appliqué sur la base de données décrite ci-dessus afin de procéder d'abord à une sélection de variables et ensuite sur les variables sélectionnées pour évaluer la pertinence des résultats. Au niveau 1 de l'algorithme, plusieurs applications de 2S-SOM ont été réalisées en faisant varier les paramètres d'initialisation. La meilleure carte en termes de quantification vectorielle est retenue et on a obtenu une carte de 21×12 soit 252 neurones. La projection des variables du paramètre BLH de la base de données sur la carte obtenue montre une bonne organisation de la topologie des observations sur la carte (Cf. FIG 1).

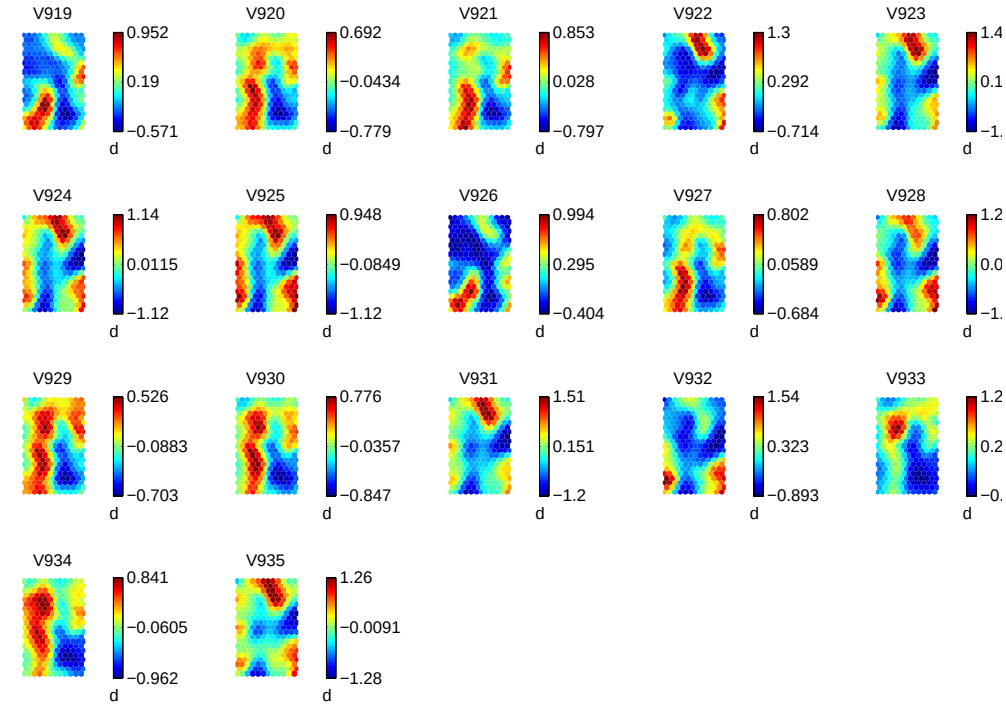


FIG. 1 – *Représentation de la topologie des variables du paramètre BLH sur la carte topologique (la barre des couleurs représente l'échelle des données, des plus faibles (bleu) aux plus fortes (rouge))*

Les autres variables n'ont pas été représentées à cause du nombre important de variables d'apprentissage, mais un travail préalable a été effectué pour s'assurer de leur bonne organisation avant de procéder à l'exploitation de cette carte topologique.
La figure 2 suivante présente le nombre de données captées par chaque neurone de la carte. Nous remarquons qu'il existe une distribution relativement homogène des données sur la carte.

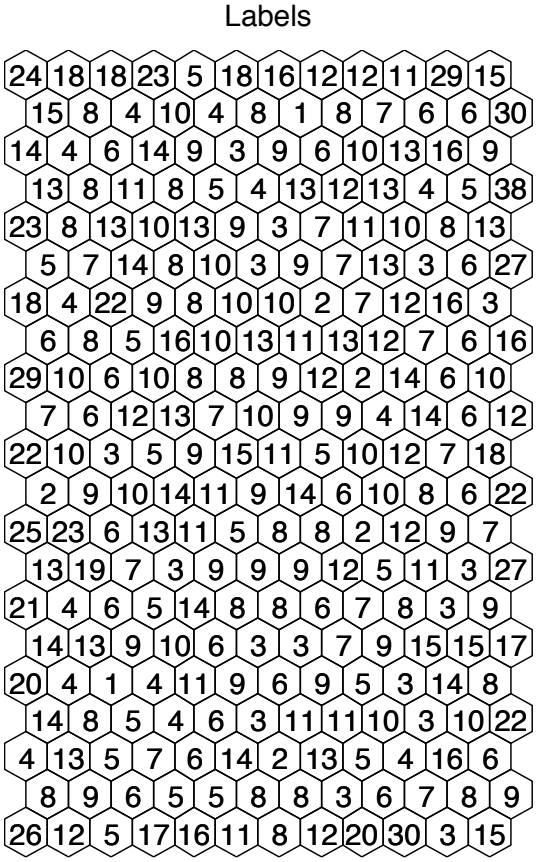
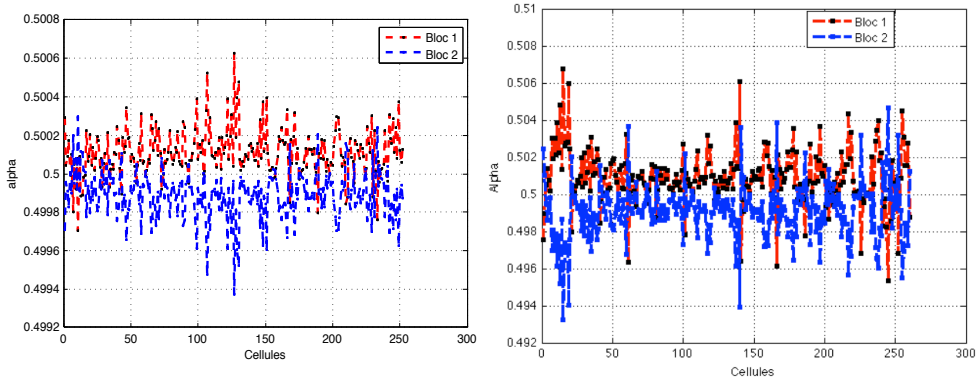


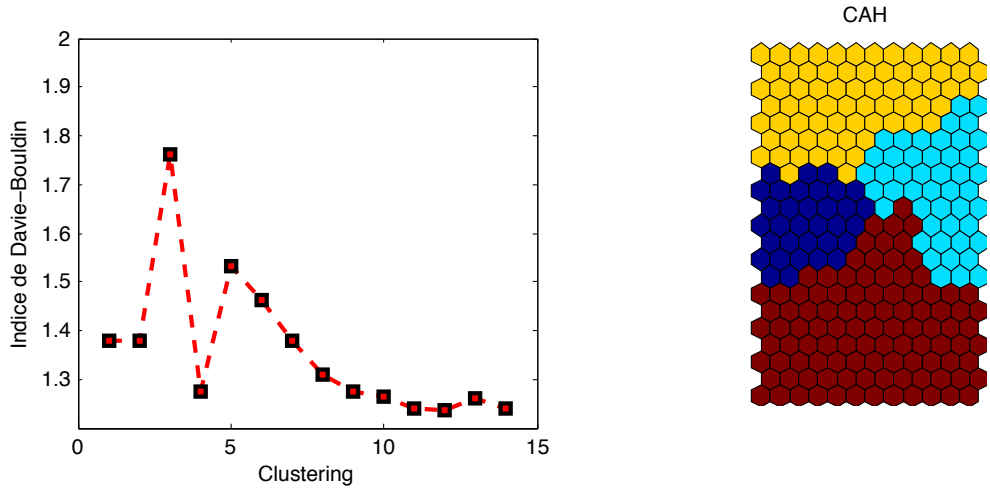
FIG. 2 – *Cardinalité des neurones de la carte topologique*

La figure 3(a) donne une représentation graphique des poids α_{cb} définis par 2S-SOM sur les blocs par rapport aux cellules de la carte sur les deux niveaux. Il ressort de l'analyse de cette dernière, que pour le niveau 1 comme au niveau 2, globalement le bloc 1 est plus important que le bloc 2 sur les cellules. Ce qui montre que la suppression des variables de bruit ne change pas fondamentalement l'importance relative des blocs.

Étant donné le grand nombre de neurones de la carte, nous avons regroupé les neurones, en appliquant une classification ascendante hiérarchique (CAH) avec le critère de Ward comme critère d'agrégation des neurones de la carte auto-organisatrice. La figure 3(b) suivante présente les indices de Davies et Bouldin (1979) pour plusieurs classifications en faisant varier le nombre de classes. Nous utilisons ce dernier pour choisir le bon nombre de classes, en l'occurrence 4 classes fig 3(b). les poids β de ces 4 classes servent ensuite à sélectionner les variables pertinentes pour déterminer la carte finale. Les figures 2 et 3(b) présentent la cardinalité des observations par cellule et la représentation des classes de la CAH. Le nombre de données de chaque classe présenté sur le tableau 3 montre une répartition homogène des données dans les classes.



(a) Les poids des blocs au niveau 1 (à droite) et au niveau 2 (à gauche) pour les cellules de la carte



(b) Indices de Davies Bouldin

FIG. 3 – *Caractéristiques des cellules et la répartition en classes*

	classe1	classe2	classe3	classe4
N	650	525	556	818
P	25.5%	20.6%	21,80%	32.09%

TAB. 3 – *nombre de données dans chaque classe et le pourcentage associé)*

La sélection des variables a été réalisée en deux étapes. Nous avons utilisé d'abord le principe des valeurs tests présenté dans la section 2.2 pour évaluer les variables pertinentes dans les 4 classes obtenues par la CAH afin de procéder à un premier filtre tenant compte de la

variabilité intra classe des poids β de chaque variable. Ensuite sur les variables retenues par la valeur test, on a sélectionné les variables les plus pertinentes en se fixant un seuil à $1/p_b$ avec p_b étant le nombre de variables du bloc b. On dira qu'une variable est importante si son poids moyen dans une classe est significativement supérieur à $1/p_b$. La figure 4, montre le poids des variables de chaque bloc dans chaque classe au niveau 1. La sélection des données effectuée montre que 59% (548 variables dont 408 pour le bloc 1 et 150 pour le bloc 2) des variables ne sont pertinentes pour aucune classe. Les 41% (377 variables dont 204 pour le bloc 1 et 173 pour le bloc 2) des variables restantes ont servi à déterminer la partition finale au niveau 2.

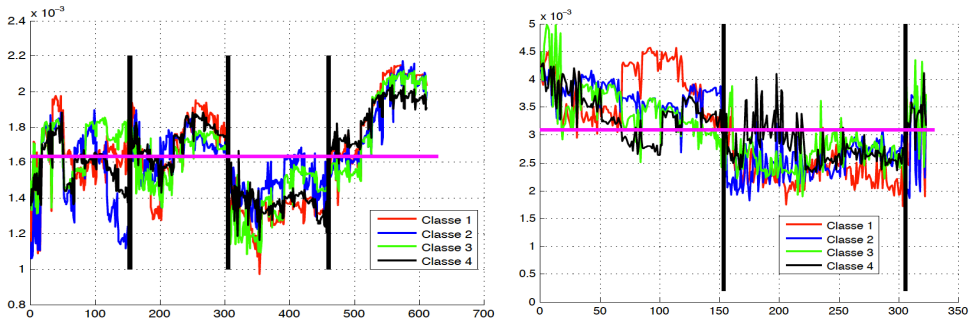


FIG. 4 – *Les poids des variables des blocs au niveau 1 (à gauche) et au niveau 2 (à droite) pour les cellules de la carte*

3.3 Description des classes

La figure 5 caractérise les 5 classes par rapport aux vents qui décrivent la direction et l'intensité du vent. On observe que les classes 1, 2, 5 sont caractérisées par des vents venant du Nord Est alors que les classes 3 et 4 sont caractérisées par les vents venant de l'Ouest.

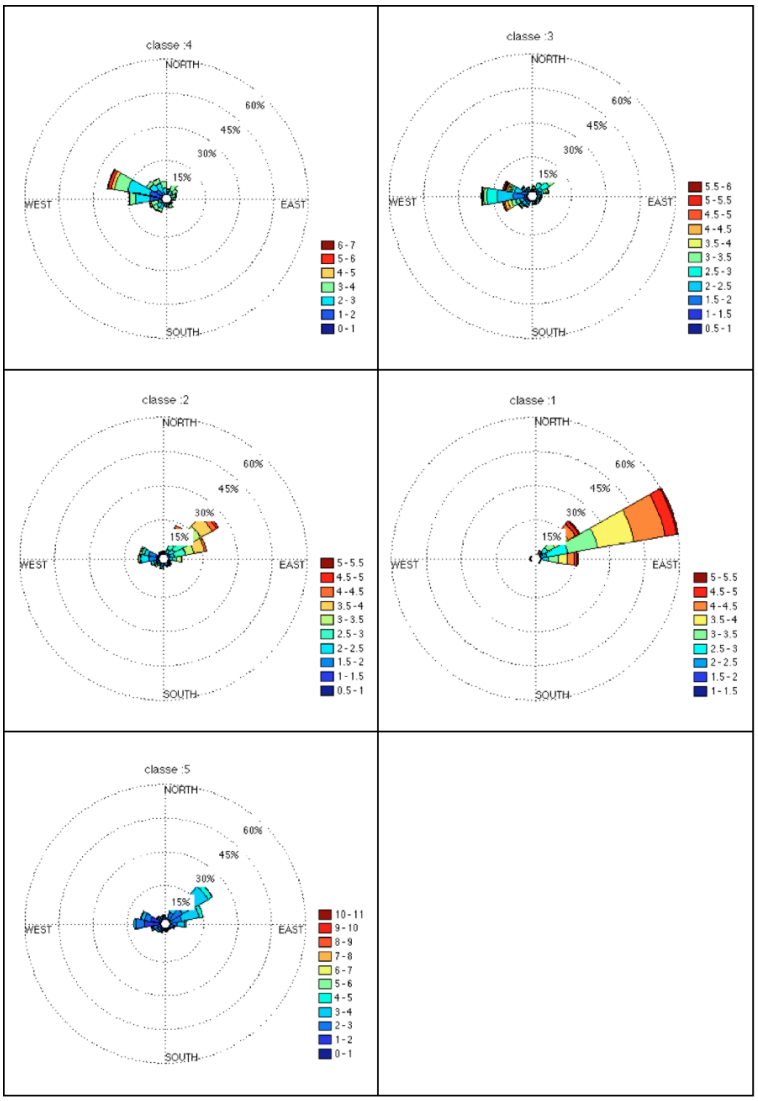


FIG. 5 – *rose des vents de chaque classe*

La figure 6, montre que les classes 1, 2 et 5 sont essentiellement constituées des données de la saison sèche (de Novembre à Mars), la classe 3 est majoritairement constituée des données du mois de Mai, mois d'intersection entre la saison sèche et la saison de pluie et la classe 4 est majoritairement constituée des données du mois d'Avril. Nous remarquons également que les classes correspondant aux mois de la saison sèche sont caractérisées par des vents venant du nord Est c'est à dire les vent d'harmattan (Figure 6). Les classes 3 et 4 correspondant majoritairement aux mois de début de la saison des pluies se caractérisent par des vents venant

majoritairement de l'ouest. Ce qui indique que l'utilisation de 2S-SOM au niveau 2 a permis d'avoir des classes avec une saisonnalité et une direction de vent particulière chacune (Figure 6, 5) et qui permet de dire qu'à la sortie, la sélection de variables permet d'avoir des résultats cohérents interprétable au point de vue géophysique.

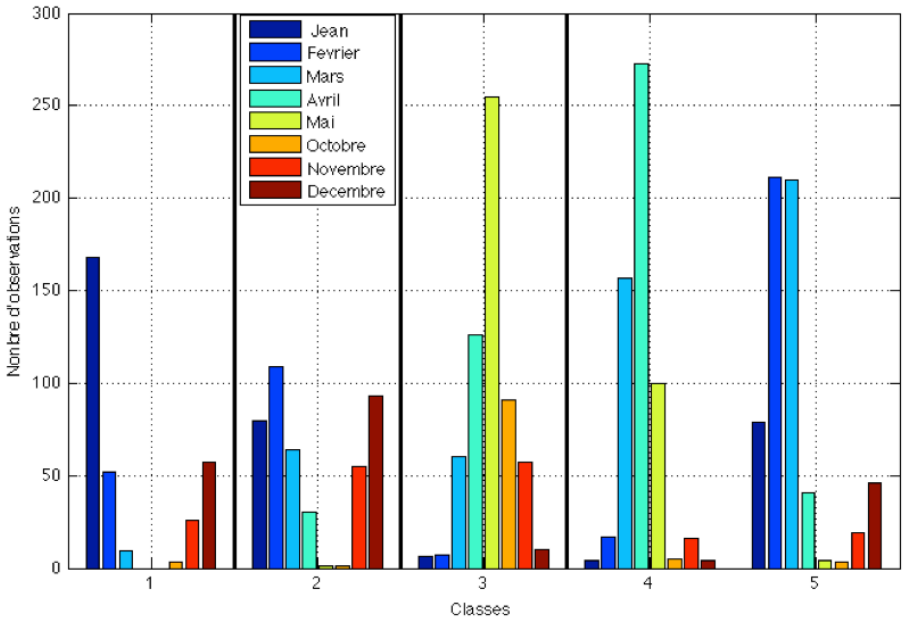


FIG. 6 – *Saisonnalité des classes obtenues au niveau 2*

4 Conclusion

Nous avons proposé une approche de sélection des variables en classification. Les meilleurs résultats de classification obtenus sur les variables pertinentes sélectionnées au niveau 1 vis-à-vis des données étiquetées montre l'intérêt de ce processus de filtrage des données. De plus, l'application de la méthode sur les données météorologiques montre que nous avons proposé une méthode efficace de sélection de variables permettant de fournir en sortie non seulement des variables pertinentes mais aussi un résultat visuel et compréhensible des clusters identifiés.

Références

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

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

Davies, D. L. et D. W. Bouldin (1979). A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (2), 224–227.

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

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.*

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

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.

Lebart, L., A. Morineau, et M. Piron (1997). Statistique exploratoire multidimensionnelle.

Ouattara, M., N. Niang, F. Badran, et C. Mandin (2014). une méthode de soft-subspace clustering pour données multi-blocs basée sur les cartes topologiques auto-organisées. *Revue des Nouvelles Technologies de l'information (RNTI).*.

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

Annexe

Les indices de comparaison de deux partitions C et C'.
Nous désignons par :

— N_{11}, le nombre de fois où deux observations sont dans une même classe dans C et dans une classe C' (accords positifs)
— N_{10}, le nombre de fois où deux observations sont dans la même classe de Cet dans des classes différentes dans C'.
— N_{01}, le nombre de fois où deux observations sont dans la même classe de C et des classes différentes C'
— N_{00}, le nombre de fois où deux observations sont dans des classes différentes de C et de C' (accords négatifs)

L'indice de précision indique la probabilité que deux objets soient regroupés dans la partition C s'ils le sont dans la partition C' :

$$Prcision(C, C') = \frac{N_{11}}{N_{11}+N_{01}}$$

Le coefficient de rappel évalue la probabilité que deux objets soient regroupés dans la partition C s'ils le sont dans la partition C' :

$$Rappel(C, C') = \frac{N_{11}}{N_{11}+N_{10}}$$

La pureté d'une partition s'évalue en quantifiant la cohérence d'une partition par rapport une autre. La manière la plus simple d'évaluer la pureté est de rechercher le label majoritaire de chaque classe et de sommer le nombre d'observations ayant le label majoritaire par classe. La pureté se définit alors simplement par l'expression suivante :

$$Purete(C, C') = 1/N \sum_{k=1}^{K} argmax_{cl}(n_{kl})$$

n_{kl} est le nombre d'observation dans la classe k de C et dans la classe l de C'.
L'indice de Rand indique la proportion de paires d'observations pour lesquelles deux partitions sont en accord.

$$Rand(C, C') = \frac{N_{11}+N_{00}}{N_{11}+N_{11}+N_{01}+N_{00}}$$

Summary

We propose a method of feature selection in classification based on self-organized maps SOM. It uses the method of subspace clustering 2S-SOM in two hierarchical steps. The first level provides a system of weight, evaluating the variables and blocks relative contributions to the groups of observations. These weights allow the selection of relevant variables. 2S-SOM is again used on the selected variables to determine the final partition of the observations in the second step. The method is evaluated on simulated and real data. In particular, the application on meteorological data shows that the selection of variables at the level 1 facilitates the geophysical interpretation of the classes obtained at the level 2.