Hierarchical Mixed Topological Maps

Ndèye Niang*, Mory Ouattara *,**

*Laboratoire CEDRIC, CNAM 292, rue Saint Martin, 75141 Paris Cedex 03, France,
n-deye.niang_keita@cnam.fr,
http://cedric.cnam.fr
**Centre Scientifique et Technique du Bâtiment, 84 Avenue Jean Jaurès, 77420
Champs-sur-Marne
mory.ouattara@cstb.fr,
http://cedric.cnam.fr

Abstract. We address the problem of clustering individuals described with several mixed variables divided in homogeneous blocks. We propose a hierarchical method with two levels to partition the individuals. The method is based on two successive steps using mixed topological maps combined with agglomerative hierarchical clustering. The proposed approach allows to take into account simultaneously qualitative and quantitative variables as well as the variable blocking. A real example on indoor air quality illustrates the proposed method.

1 Introduction

Clustering analysis is probably one of the most widely used statistical tools in data mining. In various applications in engineering, biology, business and social sciences classical clustering methods as well as new approaches are more and more used to reduce huge transactional and experimental data. These data are often column partitioned, as the variables are divided in several homogeneous and meaningful blocks. For example in biology or chemistry experimental analysis, a data block may consist of a set of measurement variables which refer to the same type of instrument or method used for the analysis; alternatively, a block may contain variables having biological similarity. In the indoor air quality study which illustrates the method we propose, several questionnaires related to different possible causes of pollution have been fulfilled separately. The resulting data sets contains a combination of categorical and continuous variables, we refer to them as mixed data sets.

The general problem addressed in this paper is discovering unknown homogenous groups of observations keeping into account the multiblock structure of the data. We propose a hierarchical approach with two levels of clustering using agglomerative hierarchical clustering combined with mixed topological maps (MTM) (Lebbah et al., 2005), a modified version of self organizing map (SOM) (Kohonen, 1982, 1995, 1998) for mixed variables. The first step clusters each initial block of variables separately providing local partitions (ie according to each block). In the second step, based on the results of the former one, the different local partitions are combined into a single global one. The aim of this approach, called hierarchical mixed topological map ($HMTM$) is interpreting block-specific patterns of heterogeneity

and providing at the same time a synthesis of the information about the clusters shared by the blocks.

The remainder of the paper is organized as follows. Section 2 recapitulates the background on SOM based clustering with mixed data, with a focus on MTM. The proposed method is described in Section 3 and exemplified on indoor air quality data in Section 4. The paper closes with a discussion in Section 5.

2 Topological maps for mixed data

The objective of cluster analysis is to discover significant groups of individuals. Clustering techniques achieve a reduction of the data size by gathering items having similar descriptions in homogeneous groups well separated and whose members are close to each other according to some proximity measure. A survey of clustering techniques can be found in Jain (2010). These methods are also briefly reviewed in Kotsiantis and Pintelas (2004) with a focus on recent methods of ensemble clustering.

The two most popular approaches to clustering are the agglomerative hierarchical clustering, that proceeds successively by merging small clusters into larger ones, and the direct partitioning methods such as the widely used K-means and SOM (Kohonen, 1982, 1995, 1998) in the framework of neural network. This paper is focussed on the partitioning techniques based on SOM.

Self Organizing Map is an unsupervised learning method that achieves both tasks of non linear projection and clustering. The genuine version of SOM only processes quantitative data. Quite recently, new models of topological maps have been proposed for clustering data with catagorical attributes as well as mixed data sets. Using a probabilistic formulation of clustering problem as a mixture modeling problem, allows extending the classical gaussian mixture for numerical variables to others distributions adapted for categorical variables such as multinomial distributions (Jollois and Nadif, 2002). In particular, extending the mixture modeling interpretation of SOM to categorical variables distributions leads to probablistic topological maps (Anouar et al., 1998; Heskes, 2001; Lebbah, 2003; Verbeek et al., 2005). This subject is not the focus of this paper and will be considered in future work.

A common feature of many methods adapted to handle mixed variables is to transform categorical variables or adapt distance measures to make it possible to analyze this type of data using techniques designed for numerical data. (Huang, 1997; Huang and Ng, 1999; Ganti et al., 1999; Guha et al., 2000). Several authors have contributed to extend SOM to categorical and mixed variables in this way (Cottrell et al., 2004; Chen and Marques, 2005; Lebbah et al., 2000, 2005).

Lebbah et al. (2000) proposed the Binbatch algorithm which is based on a binary transformation of the categorical variables and uses the Hamming distance. Then he proposed the mixed topological map method (Lebbah et al., 2005) that achieves direct clustering of units described by mixed variables by combining both SOM and Binbatch algorithms. The general principle of MTM is briefly presented below to facilitate the presentation of $HMTM$.

The basic idea of MTM, as in SOM, is to display a high dimensional data set in a space of lower (usually of one or two) dimensions. MTM consists of a set of neurons organized on a regular grid C called map. The map has a discrete topology defined by a distance $\sigma(c, r)$ equal to the length of the shortest path between c and r , a pair of neurons on the graph.

For each neuron c, this distance allows to define a neighborhood parameterized by a kernel positive function $K(lim_{(|x|\to\infty)}K(x) = 0)$ in order to control the neighborhood size. The K_T parameterized family is define by $T : K_T(\sigma) = \frac{1}{T}K(\frac{\sigma}{T})$.

Let D be the data space ($D \subset R^n$) and $A = \{z_i; i = 1, ..., N\}$ the training data set ($A \subset D$). In MTM, each observation $z_i \in A$ is considered to be in two parts: numerical part $z_i^r = (z_{i1}, ..., z_{ip})(z_i \in R^p)$ and binary part $z_i^b = (z_{i(p+1)}, ..., z_{in}) \in \{0, 1\}^{(n-p)}$. So $z_i = (z_{ir}, z_{ib})$ is a mixed vector of dimension n. As in the self organizing map, each cell c of the grid is associated to a referent vector randomly choosen from the training data set. It is a mixed vector $w_c = (w_{cr}; w_{cb})$, where $w_{cr} \in R^p$ and $w_{cb} \in \{0, 1\}^{(n-p)}$. The set of the referent vectors W is decomposed into W^r and W^b, respectively the set of numerical and binary part of the referent vectors. It fully determines the MTM and has to be estimated from A. This is done iteratively by minimizing a cost function.The MTM algorithm is derived from the batch versions of the Kohonen algorithm for numerical data and the Binbatch algorithm for binary data. In this algorithm, the similarity measure and the estimation of the referent vectors are specific to each part of the data: it is the Euclidean distance d_{Eucl} with the mean vector for the numerical data and the Hamming distance d_H with the median center for the binary data. To ensure all numerical variables have equal influence on distance, they are usually normalized. Furthermore to avoid favouring either type of variable, the Hamming distance is weighted using a fixed parameter β. Huang (1997) discusses the effect of this weight choice on the clustering process. When data are normalized in order to lie in an $[0, 1]$ interval, choosing this parameter as the ratio between the number of continuous variables and the number of binary variables is equivalent to normalize each distance to its maximal value. Rogovschi et al. (2011) propose adaptative weights which have to be estimated iteratively in a supplementary step of the MTM process. This issue of adaptive weighting is not tackled in this paper.

The MTM cost function is:

$$J^T_{MTM}(Z, \omega) = \sum_{Z_i \in A} \sum_{c \in C} K_T(\sigma(X(z_i), c))(d_{Eucl}(z_{ir}, \omega_{cr}) + \beta d_H(z_{ib}, \omega_{cb})) \qquad (1)$$

$J^T_{MTM}(Z, \omega) = J^T_{SOM}(Z, \omega) + \beta J^T_{Bin}(Z, \omega)$ where $J_{SOM}(X, W)$ is the SOM cost function, and $J_{bin}(X, W) = \sum_{Z_i \in A} \sum_{c \in C} K_T(\sigma(X(z_i), c))d_H(z_{ib}, \omega_{cb})$ is the cost function used in the Binbatch algorithm, here weighted by $\beta = \frac{p}{(n-p)}$

The minimization of the MTM cost function is made using batch iterative process with two steps as in the SOM algorithm (Anouar et al., 1998; Luttrell, 1994) which can be expressed as a dynamic cluster method (Diday and Simon, 1976).

– Assignment step : assuming that W is fixed, J_{MTM} has to be minimized with respect to X. This leads to the following assignment function:

$$X(z_i) = argmin_{c \in C}||z_i - \omega_c||^2 = argmin_{c \in C}(d_{Eucl}(z_{ir}, \omega_{cr}) + d_H(z_{ib}, \omega_{cb})) \qquad (2)$$

– Minimization step : assuming that X is fixed, this step minimizes $J_{MTM}(X, W)$ with respect to W in the space $R^p \times \{0, 1\}^{n-p}$. The minimization of the cost function (1) leads to minimize the function $J_{SOM}(X, W)$ in R^p and $J_{bin}(X, W)$ in $\{0, 1\}^{(n-p)}$. These two minimizations allow to define the numerical part w_{cr} as in SOM and the binary part w_{cb} of the referent vector w_c as the median center of the binary part of

the observations $z_i \in A$ weighted by $K^T(\sigma(X(z_i), c))$. Each component of $w_c = (w_{cr}; w_{cb})$ is then computed as follows:
for the quantitative part:

$$\omega_{cr} = \frac{\sum_{r \in C} K^T(\sigma(X(z_i), c)), z_i}{\sum_{r \in C} K^T(\sigma(X(z_i), c))} \tag{3}$$

for the binary part:

$$w_{cbk} = \begin{cases} 0 \quad if \sum_{z_i \in A} K^T(\sigma(X(z_i), c))(1 - z_{ib}^k) > \sum_{z_i \in A} K^T(\sigma(X(z_i), c)) z_{ib}^k \\ 1 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad otherwise \end{cases} \tag{4}$$

The nature of the topological model reached at the end of the algorithm, the quality of the clustering and the topological order induced by the graph greatly depend on the neighborhood function K. In practice, a parametrized function K^T is used to control the size of the neighborhood $K^T(\sigma(r, c)) = \exp(\frac{\sigma(r,c)}{T})$. For a given value of T , the batch algorithm leads to local minimum of the cost function with respect to X and W (Anouar et al., 1998). To avoid local mimima several runs of the algorithm with different starting points are done and the best partition is choosen. The previous iterations are repeated with T decreasing from an initial value T^{max} to a final value T^{min}.
At the end of the learning algorithm, the minimized cost function can be seen as a K-means one, but MTM takes into account the topological constraints. MTM inherits the SOM properties, such as the property of topology conservation and robustness which are the main characteristic of SOM compared to other clustering methods like K-means. As in SOM the individuals are put on the map while keeping the neighborhood in such way that similar individuals in the original space will be close on the map. An outlier affects only one referent and its neighborhood, thus it can be easily detected from the map since its distance in the input space from other units is large. This property can be useful in real life application where outliers may exist. Additionally, there is a difference from a pratical point of view: in the K-means clustering the number of clusters should be choosen according to the number of clusters there are in the data (which is unknown), in the SOM the number of reference vectors can be choosen to be larger, not necessary respective of the number of clusters (Kaski, 1997). It has to be noticed that the referent vectors obtained at the end of the MTM process, share the same code with the initial observations and then can be decoded in the same way, allowing a direct interpretation of the binary part of the referent vectors.

There are other methods that extend SOM to categorical and mixed variables. Cottrell et al. (2004) present in a unified way several SOM based methods (Cottrell and Rousset, 1997; Cottrell et al., 1998). They consist in applying the SOM algorithm on transformed data: Burt matrix in $KMCA$ or complete disjunctive table for $KMCA_{ind}$ and $KDISJ$ and the chi square distance (or Euclidian distance on corrected Burt or disjunctive tables) is used in both cases. For mixed variables, they proposed additional preprocessings of the data using SOM followed by AHC on the quantitative variables to define a new qualitative variable whose categories are the labels of the AHC cluster. Then $KMCA$ is used on the new table obtained by adding the new cluster variable to the other (original) qualitative variables. This does not allow

direct interpretation of the referent vector as pointed in (Lebbah, 2003). They also propose to transform qualitative variables into numerical ones by a multiple correspondance analysis (MCA) and then keeping all the factorial coordinates it is possible to use the classical SOM. Chen and Marques (2005) propose $NCSOM$ to tackle the mixed variable problem using a combination of the mismatch measurement for the categorical features and euclidian distance equally weighted. $NCSOM$ add to the SOM process an additional step based on a threshold which has to be fixed.

It could be interesting to conduct a comparative study of these approaches but it is not the purpose of this paper.

3 Hierarchical Mixed Topological Maps

Hierarchical methods such as $HPCA$ and $CPCA$ (Westerhuis et al., 1998) have been proposed to analyze data sets with very large number of variables which can be structured in conceptually meaningful blocks. They are used in order to improve interpretability of multivariate analysis of a data set when the results are difficult to interpret due to the huge number variables. In $HPCA$, the variable blocking is taken into account and that leads to a two level model: the local or lower level, in which the standard method (PCA) is performed on each block separately, and the upper or global level, where the relationship between blocks are modeled by running again the standard method on the block scores. The principles of hierarchical and multiblock PCA is reviewed in Wold et al. (1996) and reconsidered in Westerhuis et al. (1998) from a theoretical and an algorithmic point of view.They are aimed at summarizing relationship between variables rather than highlighting heterogeneity between individuals.

For clustering on multiblock mixed data, we propose to extend the hierarchical principle of Wold's $HPCA$ to MTM. Thus, $HMTM$ consists of two successive applications of MTM: first to each of the initial blocks of variables separately and in a second step on the results of the former step. The method is detailed below.

3.1 Notations

The following notations are adopted in the rest of the paper:
- N : number of individuals indexed $i = 1, ..., N$
- B : number of blocks indexed $b = 1, ..., B$
- q_b : number of variables in block b
- Z_b : an $N \times q_b$ matrix containing the observed values of N individuals on q_b variables
- z_{bi} : an $1 \times q_b$ vector containing values of individual i in block b
- p_b : partition of block b into k_b clusters
- b_c : cell c of the block b
- w_{bc} : referent vectors of the cell(or cluster) c of the block b
- W_b : an matrix $N \times q_b$ in which each individual i is represented by its assigned referent vector
- V : an $N \times (q_1 + ... + q_B)$ matrix obtained by horizontal merging of W_b
- Ω : space referents at level 2 of $HMTM$
- Ω_c : referent vector of cell c in the map associate of table V

3.2 HMTM algorithm

The first step of $HMTM$ consists of B separated applications of MTM to each data set Z_b . Each MTM process has its own set of parameters, no constraints on the number of referents as in some consensus methods. This will be discussed later. Through the iterative process described in section 2 the cost function to minimize is $J_{MTM}(X, W)$ defined in relation (2).

Thus the first part of the $HMTM$ method will provide B maps defining the B best partitions p_b. In order to facilitate the cluster interpretation task, we applied an Agglomerative Hierarchical Clustering (AHC)to reduce the number of clusters. This will be illustrated in the case study in section 4.

Starting from the partitions p_b, we propose to represent each of the N initial observations by its assigned referent vector w_{bc} corresponding to the cell b_c (or cluster) to which the observation belongs in the MTM map of block b. This prototype representation yields B new data sets, each denoted W_b, summarizing the local structure of the individuals according to the variables of block b. Afterwards these data sets are horizontally merged in a unique table V which is used as the input for the second level of $HMTM$ (Fig.1). More precisely, in the second step, each of the N initial observations is described by $v_i = (w_{1ci}, ..., w_{Bci})$. The new matrix V is then used to perform the second MTM algorithm with its own set of parameters. So $HMTM$ is an adaptation of the hierarchical principal component analysis (Wold et al., 1996) to clustering with few slight differences as it will be discussed below. The whole $HMTM$ process is presented in Fig.1 below.

The cost function of the second step is then J^T_{HMTM} and the iterative process yields the following solutions for the qualitative and quantitative part of the referent vectors :

$$J^T_{HMTM}(Z, \Omega) = \sum_{V_i \in A} \sum_{c \in C} K_T(\sigma(X(V_i), c))(d_{Eucl}(V_i^r, \Omega_c^r) + d_H(V_i^b, \Omega_c^b)) \tag{5}$$

$$\Omega_c^r = \frac{\sum_{r \in C} K^T(\sigma(X(V_i), c))V_i}{\sum_{r \in C} K^T(\sigma(X(V_i), c))} \tag{6}$$

for binary part

$$\Omega_c^{bk} = \begin{cases} 0 & if \sum_{V_i \in A} K^T(\sigma(X(V_i), c))(1 - V_i)^{bk} > \sum_{V_i \in A} K^T(\sigma(X(V_i), c))V_i^{bk} \\ 1 & otherwise \end{cases} \tag{7}$$

As for the maps of the first level, an agglomerative hierarchical clustering can be performed from the second level map to obtain a reduced number of easier interpretable clusters. While being ignored in derivation of local partitions, relationships among variables of different blocks are taken into account at the second level. The summarized representation can be seen as a way of fusing the blocks after having reduced the block information using the referent vectors. The resulting global clustering yields a combination of local partitions.

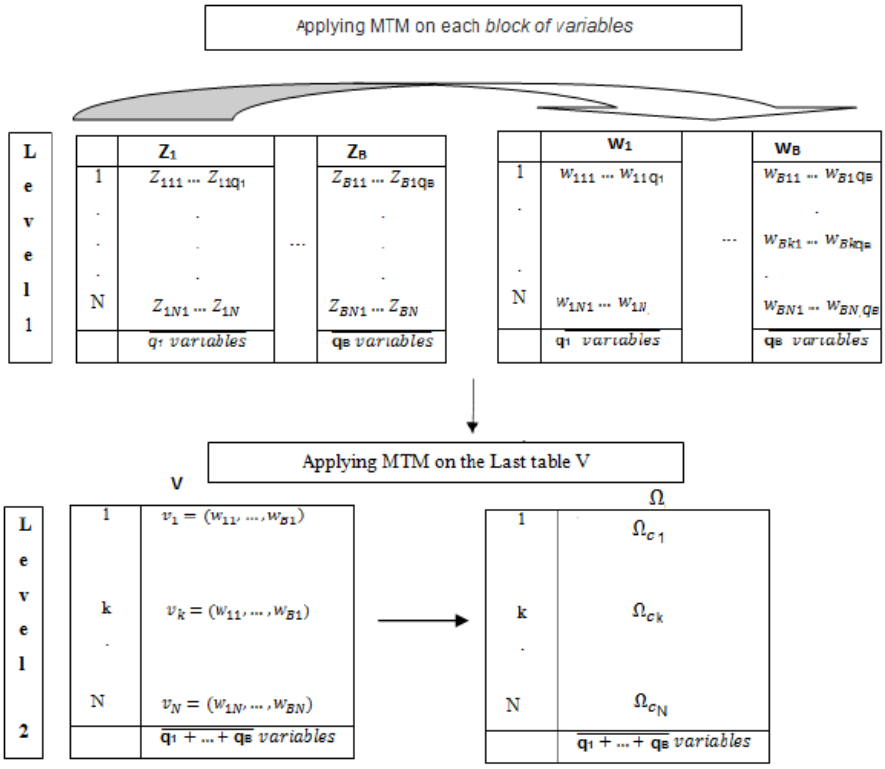


FIG. 1 – *The HMTM process*

3.3 Cluster quality and interpretation

As clusters to discover are unknown a priori, the final partitions of a data set require some kind of quantitative evaluation in most of applications (Halkidi et al., 2001). Several relative criteria have been proposed. In an evaluation study of thirty validity indices (Milligan and Cooper, 1985) it is noted that the results are likely to be data dependent.

To assess the map quality we use the basic Davies-Bouldin index DB_K (Davies and Bouldin, 1979) for its simplicity and interpretability in terms of average dispersion. Other validity indices like the silhouette-value (Rousseeuw, 1987) or the Hubert statistic (Hubert and Arabie, 1985; Arabie and Hubert, 1994) can also be used. The DB_K index (Davies and Bouldin, 1979) is defined as:

$$DB_k = \frac{1}{K}\sum_{k=1}^{K}\max_{i\neq j}\frac{R_i + R_j}{R_{ij}} \tag{8}$$

where K is the number of clusters, R_i is the average distance of all individuals in the cluster C_i to their cluster center c_i and R_{ij} is the distance between the cluster centers c_i and c_j . Hence, small values of Davies-Bouldin involve that clusters are compact and far from each other.

Other choices of the measure of dispersion of a cluster and the dissimilarity between two clusters yield different DB_K index (Davies and Bouldin, 1979) . The DB_K index in the above definition is the average similarity between each cluster and its most similar one. The DB_K index exhibits no trends with respect to the number of clusters and thus the minimum value of DB_K in its plot versus the number of clusters will provide the best partition.

For interpreting clusters, we determine the most important variables characterizing the partition through a Wilcoxon test (Hollander, 1973) for numerical attributes and through a comparison test of proportion for categorical attributes (Lebart et al., 2006).

4 Application to indoor air quality data

Now we present an application of $HMTM$ to illustrate the effect of taking into account the block structure of the variables on the quality of the partitions and on the interpretation of the clusters.

4.1 Data presentation

The data was collected in the framework of the national survey carried out by the Indoor Air Quality Observatory (IAQO) from October 2003 to December 2005 (Kirchner et al., 2009). The study consisted in a cross-sectional survey in sample of dwellings drawn from the entire sample of all principal residences in mainland France (24 millions). A three-stage random selection procedure was used to obtain a representative sample of dwellings taking into account the municipalities proportion to their number of main residences, the land registry sections within municipalities and main residences within the land registry section. The final sample included 567 main residences distributed among 74 municipalities of 55 departments and 19 regions. Different categories of questionnaires were used in addition to the measurement of key pollutants and parameters selected on the basis of their potential impact on air quality. A face-to-face household questionnaire described outdoor environment, building characteristics, equipments, and occupants living and cleaning habits. More than 650 variables were collected per dwelling. After preliminary studies 125 mixed variables were extracted from the initial data set. These variables were divided into three meaningful blocks describing respectively: the structure of the dwelling (32 quantitative variables and 39 qualitative variables), the characteristics of the household (5 quantitative variables and 6 qualitative variables) and the living habits of the inhabitants (21 quantitative variables and 23 qualitative variables). The block of pollutants attributes not considered in this application will be used in secondary analysis to study potential links between the pollutants and the dwellings characteristics.

4.2 Application of HMTM

$HMTM$ was applied to the 125 mixed variables structured in three blocks. This provides three structures of the dwellings according to the separated blocks and their combined partition

and gives their cluster characterization with the variables used to get the partitions. The MTM learning process has been performed separately several times on blocks of variables by varying the parameters of the cost function, the number of iterations and the map dimensions. The basic Davies Bouldin-index was choosen for cluster validity studies giving then the best map for each block. A map of 8×8 neurons, corresponding to 64 clusters, is obtained for the structure of the dwelling block. The habits of inhabitants block gives a 9×9 map (81 clusters) and the block of characteristics of the household yields a map with 6×6 neurons (36 clusters). Then the three partitions results are merged in the summarized data set denoted V which is used afterwards to perform the second MTM learning process. The final global partition is obtained with a map of 7×7 neurons (49 clusters). The numbers of clusters provided by local partitions as well as the global one are relative large. So starting with the three local maps corresponding to the structure of the dwelling, the characteristics of the household and the living habits of the inhabitants we apply an agglomerative hierarchical clustering which yields final local partitions of dwellings with reduced number of clusters respectively equal to 5, 6, and 8. Figure 2 shows the map on which the final local partition is visualized only for the structure of the dwellings block for shortness. The Davies-Bouldin index evolutions for the MTM process and for the hierarchical clustering are also presented. Its values are 0.95, 0.92, and 1.15 respectively for the block of the structure of the dwellings, the characteristics of the household and the habits of the inhabitants. As the Davie-Bouldin index is independent of the cluster number, it can be considered as a quality indicator of different clusterings of the blocks in $HMTM$ level 1. After the AHC the final local partitions allow an easier characterization of groups of dwellings focused on variables specific to each block. The 49 clusters of the global map have been also reduced by AHC into 8 groups interpreted using all the variables figure 2(b). The detailed interpretations of the partitions have been done using the topological map visualization capabalities and the statistical tests cited in subsection 3.3. TAB-2 presents the list of main significant variables. For this present application, we are interessed in comparing the partitions and highligthing the main impacts on significant variables for the clusters, in order to illustrate the effect of taking into account the block structure. Local partitions from the level 1, the final level 2 partition of $HMTM$ and a direct partition on the unblocked data are compared in terms of Rand index.

The C code used to implement the algorithms for the mixed topological maps is available at https://docs.google.com/folder/d/0B0lRULTv1q7XblpRVVY1UHNMWlE/edit.

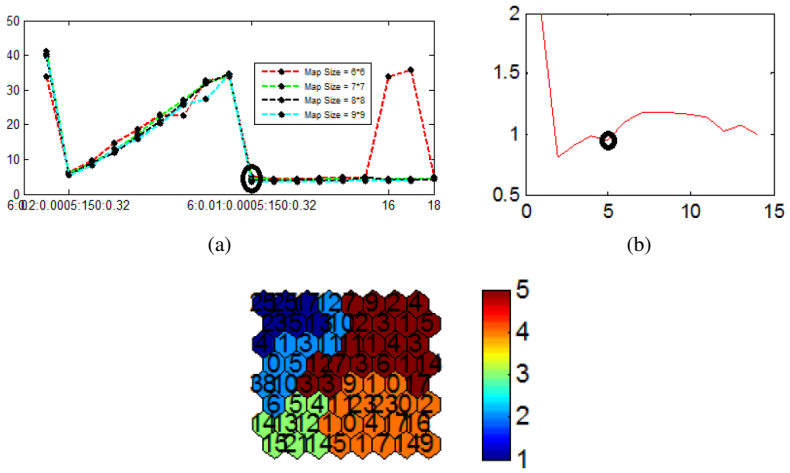


FIG. 1 – *The choice of the best clustering: (a) the Davies-Bouldin index for different initializations of parameters learning; (b)The final clusters obtained after level 1 of the HMTM followed by a Agglomerative Hierarchical Clustering for the block of structure of dwellings; and, (c) the final map;*

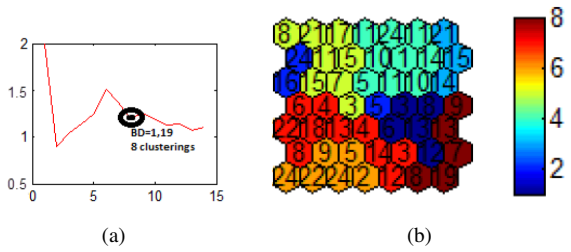
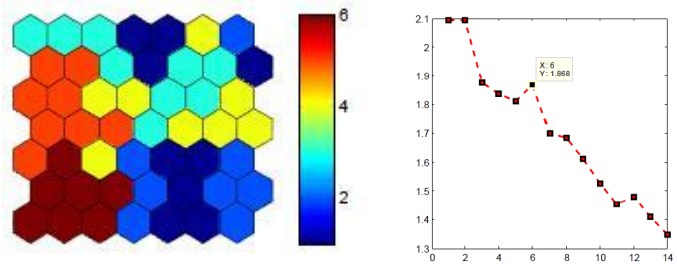


FIG. 2 – *The choice of the best clustering: (a) The final clusters obtained after level 2 of the HMTM followed by a Agglomerative Hierarchical Clustering and (b) the final map;*



(a) (b)

FIG. 3 – *(a) The final clusters of the MTM_{gen} followed by a Agglomerative Hierarchical Clustering and (b) Davies-Bouldin index;*

Cluster interpretation

In the block of the structure of the dwelling, a local partition of the individuals into five clusters is obtained. The most characteristic variables of these clusters are: the surface of the dwelling (HSRF), the age of the building (NIACE) and the furniture inside the home (wall and soil surface treatment, agglomerate and solid woods, wood furniture rate (TMM1, TMMB1, TMMB2, TRS2,...). For example, cluster 1 contains the recent small collective houses (HSRF-, NIACE+). Cluster 4 includes the big individual houses (HSRF+) with a lot of combustion appliances connected to a flue (NSQ31+). TAB-2 details the results of the partitions and the list of most significant variables characterizing the clusters. It is also possible to use the variables of the other blocks as supplementary variables to enrich the cluster interpretation. We found that only two variables from the block of habits of the inhabitants related to the house maintaining (ACT) and personal care habits (QPPB2) and few variables of the household block (income (REV), age of housholder (AGE), size of the family (NPB), young children (NBPEI10)) can be used to significantly characterize the clusters of the partition of the block of the dwellings structure.

According to DB_k index, household block provides the best clustering structure. In its local partition into six clusters, the most characteristic variables for the clusters are: household income (REV), the family size (NBP), age of housholder (AGE). Using supplementary variables from the other blocks we found previously significant variables (TRS2, TMM1, NBPEI10) which are no more significant for the interpretation of the partition (see TAB-2). Thus, from a partition to another, there is a loss of the variable information in the interpretation of the clusters : a variable which significantly distinguishes two clusters of one partition does not do the same for the other partition when it is used as supplementary variable an vice versa. Similar problems in the interpretation affect the block of habits of inhabitants.

This short comparison of list of significant variables allows to see that it is not very reasonable to choose the best local partition as the one corresponding to the structure of the individuals and use the variables of other blocks to interpret it. To obtain a more interpretable characterization of the clusters using all the available information carried by the whole set of variables the global analysis at level 2 is necessary.

This second step analysis provides a final global partition into eight clusters. The most significant variables in the interpretation of the clusters are the same as those of the local partitions: REV, NBP, AGE, NBPEI10, TRS2, TMM1, TMMB1 among others. Moreover, new significant variables appear unlike in the local partition of each block separately: children older than ten years (NBPSE10) or ICOS3, QME1, TMM3 for example. The description obtained at the level 2 combines variables of each block highlighting the relationship between them (see TAB-2).

Partition comparison

The simplest way for combining the variables consists in fusing the blocks directly without reduction using referent vector. Applying MTM on the unblocked data set directly provides a single partition (MTM_{gen}). Notice that this would not allow a detailed study of individuals only based on the variables belonging to the same block. Table 2 contains the Rand indices to compare this partition and those provided by the two levels of $HMTM$ to each other. The highest values correspond to $HMTM$, that means $HMTM$ global partition has a higher level of agreement with the local partitions than the MTM_{gen} partition. Furthermore $HMTM$ realizes the best consensus among the local partitions. Another feature in this table is the quite important values (around 0.70) of the Rand indices between local partitions. This indicates that the blocks share an important part of the information underlying the structure of the individuals which has been captured by level 1 clustering. Despite this high similarity with each other, the blocks of variables have their own part of specific information which the level two analysis attempts to capture quite successfully.

Figures 1(c), 2(b) and 3(a) show the topological maps provided by the MTM on a block (the “Dwellings" block is used as an example), $HMTM$ and MTM_{gen} respectively. As one can expect, specific block partitions are more homogeneous (in “Dwellings" block case, DB index equals 0.95); $HMTM$ is able to keep a quite well structured map (DB= 1.19), whereas MTM_{gen} yields a poorly structured map (DB= 1.87).

	Dwelling	Household	Habits	HMTM	MTM_{gen}
Dwelling	1	0.69	0.70	**0.82**	0.73
Household		1	**0.68**	**0.68**	0.67
Habits			1	**0.72**	0.66
HMTM				1	**0.70**
MTM_{gen}					1

TAB. 1 – *Rand indices for comparing partitions: Dewlling, Household, Habits, HMTM, MTM_{gen} correspond to partitions obtained from dwellings blocks, household block, habits block, level 2 of HMTM and finally the partition obtained on the data set of unblocked variables*

Block 1 Dwellings	Active variables	Supplementary Household	Supplementary Habits
Classe 1 (86 obs)	HSRF-, NIACE+, TMMB2+, TRS2+	NBP	ACT
Classe 2 (197 obs)	HSRF-, TMMB2+, TRS2+	REV,	QPPB2
Classe 3 (75 obs)	NIACE+, TMMB1+, TRS2-	AGE, NBP	
Classe 4 (72)	HSRF+, TMMB1+, NSQ31+,	AGE, NBPEI10	
Classe 5 (137)	HSRF+, TMM1+	AGE, NBP	
Variables	HSRF, NIACE, TMM1, TMMB1, TMMB2, TRS2, NSQ31	NBP, NBPEI10, REV, AGE	ACT, QPPB2

Block 2 Household	Supplementary Dwellings	Active variables	Supplementary Habits
Classe 1 (56 obs)	NSQ31	AGE+, REV-, NBP-	
Classe 2 (37 obs)	HSRF, NSQ31	NBP-	ACT, QPPB2
Classe 3 (187 obs)	NSQ31, TMMB2	REV+, NBP+,	ACT, CUI1
Classe 4 (115 obs)	HSRF, NSQ31, TMMB1	AGE-, NBP+	ACT, CUI1
Classe 5 (106 obs)	HSRF, NIACE	AGE-, REV-, NBP-	ACT, CUI1
Classe 6 (66 obs)		AGE-, REV+, NBP+,	CUI1
Variables	HSRF, NSQ31, TMMB2, NIACE, TMMB1	AGE, REV, NBP	ACT, QPPB2, CUI1

Block 3 Habits	Supplementary Dwellings	Supplementary Household	Active variables
Classe 1 (65 obs)		NBP-	QPD2b+,
Classe 2 (150 obs)	HSRF	AGE, NBP, REV	CUI1+, QME2+
Classe 3 (187 obs)	HSRF	AGE, NBP, REV	ACT+, QPPB2+
Classe 4 (187)		NBP	ACT+, CUI1+
Classe 5 (82 obs)			ACT-, CUI1-
Classe 6 (129 obs)		NBP	ACT-
Classe 7 (38 obs)	HSRF	AGE, NBP, REV	ACT+, CUI1+, QPD2b-
Variables	HSRF	NBP, REV, AGE	ACT, QPPB2, CUI1, QPD2b, QME2

HMTM	Dwellings	Household	Habits
Classe 1 (42 obs)	TMM1+, HSRF+, NSQ31-	NBEI10+	ICOS3-, QPE1b-,
Classe 2 (40 obs)	TMMB1+, TMM3+, TMMB2-, TRS2-	AGE+, NBP-	CUI1+, QPD2b+, QME2+, ICOS3-
Classe 3 (55 obs)	TMMB1+, HSRF+, NSQ31+, TRS2-, TMMB2-	NBP+, REV+, NBPES10+	ACT+, CUI1+, QME2+
Classe 4 (97 obs)	TMM1+, TMMB1+, HSRF+, NSQ31-, NIACE-, TRS2-, TMMB2-	REV+, AGE+	QME1-
Classe 5 (87 obs)	TMM1-, HSRF-, NIACE+	NBES10+, NBEI10+, NBP+, REV+, AGE-	ACT+, CUI1+
Classe 6 (86 obs)	TMM1-, TMMB1-, HSRF-, NSQ31-, NIACE+, TMMB2+, TRS2+	REV+, NBES10-	QPPB2-
Classe 7 (104 obs)	TMMB1-, HSRF-, NSQ31-, NIACE-, TRS2+, TMMB2+	AGE-, REV-, NBP-, NBES10-	CUI1, ACT-, QME1-, QPPB2-
Classe 8 (56 obs)	TMM1+, NSQ31-	NBES10+, NBP+, AGE-	ACT+, QME1+, QPD2b+
Variables	HSRF, NSQ31, TMMB1, TMMB2, TMM1, TMM3, NIACE, TRS2	NBP, REV, AGE, NBPES10, NBPEI10	ACT, QPPB2, QME1, QME2, QPD2b, CUI1, QPE1b, ICOS3

TAB. 2 – *Description of each cluster obtain by MTM applying on the blocks of Dwellings, Household and Habits, and description of cluster obtain by HMTM ; (+) is high value and (-) is small value of variables on the cluster*

5 Alternative solutions

$HMTM$ is a two level clustering method which realizes a consensus of block-specific partitions. In this section we discuss briefly alternative methods for clustering block structured data and possible extensions of the method.

5.1 Consensus methods

Clustering multiblock data has been addressed by several consensus methods proposed by authors such as (Green et al., 1993; Vichi, 1998, 1999) among others. A survey on these methods can be found in Day (1986). The principal idea of these consensus methods is to agglomerate the separate partitions obtain from each block into a global partition which has to be the most similar to the contributory partitions according to some index like the Rand index. Only the labels of the clusters are used through a categorical variable unlike in $HMTM$. Furthermore, these methods usually require that the local partitions have the same number of clusters which have to fixed a priori. This constraint does not exist for consensus methods based on agglomerative hierarchical clustering as in PRINcipal CLassification Analysis ($PRINCLA$) (Vichi, 1998) which searches for consensus of dendrogram rather than consensus of partitions as in Green et al. (1993). Other more recent methods in the ensemble clustering approach (Johansson et al., 2008) also addressed the multiblock clustering. A common feature of all these methods is that they do not really handle blocks of mixed variables. They are either designed exclusively for numerical variables or they split each mixed data set into two data sets separating numerical and categorical variables loosing the meaningful concept of the block.

5.2 Tandem clustering

Factorial analysis of mixed variables can be performed through methods such as categorical principal component analysis and optimal scaling methods in general (Tenenhaus and Young, 1985) or multiple factorial analysis (MFA) (Escofier and Pagès, 1994), respectively dedicated to the analysis of one table or multiple data sets. So they can be used in the two-step approach called tandem clustering which consists in clustering on optimal components provided by a factorial method. The problem here is how to choose these components. Many warnings have been made against this two-step approach (Hubert and Arabie, 1985; Arabie and Hubert, 1994) because factorial analysis may identify dimensions that do not necessarily contribute to detect the clustering structure in the data and may obscure the recovery of underlying cluster structure. To overcome these drawbacks several methods have been proposed such as Reduced-K-Means RKM (De Soete and Carroll, 1994) and Factorial-K-Means FKM (Vichi and Kiers, 2001) These alternatives to the widely used tandem clustering are reconsidered and compared both theoretically and empirically in (Timmerman et al., 2010). These clustering methods are aimed at simultaneously achieving a clustering of the individuals and a dimension reduction of the variables. An alternative to $HMTM$ could be to extend RKM or FKM to mixed data and use these extensions instead of MTM in the fisrt level of $HMTM$. However the visualization capabilities of the topological maps will be lost. Direct clustering methods on mixed variables such as Mixed Topological Map MTM are a better alternative.

None of the listed alternative methods can be used alone to efficiently address the two issues tackled simultaneously in $HMTM$ which are the two level clustering and the mixture

of continuous and categorical variables in each block. The advantage of the hierarchical mixed topological map is it directly clusters individuals described by mixed variables while projecting thus there is not any dimension choice nor a number of clusters to fixe a priori since each map has its own set of parameters.

5.3 Weighted clustering

The proposed method is based on an extension of the principle of hierarchical PCA to clustering. However, in the $HMTM$ method presented here there isn't any block weighting unlike for Wold's method in which a mild weighting according to each block size is introduced to avoid the blocks with larger number of variables having predominant importance. The Davies-Bouldin index which can be considered as a partition quality index can be used to weight the synthetic information coming from each block of level one before using it in the learning process of level two. They can be used to consolidate the level one results that is to forced units gathered in level 1 to stay together or to relax the gathering in level 1 clusters in order to capture new relations which eventually appear when merging the results. It may be also interesting to consider adaptative weights on variables which have to be estimated iteratively as in the weighted K-Means (Jing et al., 2004). This integration of a weighting system could improve the $HMTM$ method by providing both a variable selection based on the resulting weights (variables with the lowest weights could be discarded) and a true consensus partition that is a global partition obtained by optimizing an explicitly weighted function of the local partitions. Moreover, the determination of the most important variables which is inherent to the estimation of the weights will facilitate the interpretation of the clusters.

Work is in progress on supplementary applications on other real data sets and on simulated ones to complete the evaluation of $HMTM$.

Acknowledgment

This study is part of a Ph.D. thesis in CNAM supported by French Environment and Energy Management Agency (ADEME) and Indoor Air Quality Observatory (IAQO). The authors are grateful to anonymous referees for their valuable suggestion and helpful comments that have greatly improved this paper. We thank Giorgio Russolillo, Corinne Mandin, Sylvie Thiria and Fouad Badran.

References

Anouar, F., F. Badran, and S. Thiria (1998). Probabilistic self-organizing map and radial basis function networks. *Neurocomputing 20*, 83–96.

Arabie, P. and L. Hubert (1994). *Cluster Analysis in Marketing Research. In Advanced methods in marketing research*. Oxford.

Chen, N. and N. Marques (2005). An extension of self-organizing maps to categorical data. In C. Bento, A. Cardoso, and G. Dias (Eds.), *Progress in Artificial Intelligence*, Volume 3808 of *Lecture Notes in Computer Science*, pp. 304–313. Springer Berlin / Heidelberg.

Cottrell, M., J. Fort, and G. Pagès (1998). Theoretical aspects of the som algorithm. *Neurocomputing 21*(1-3), 119 – 138.

Cottrell, M., S. Ibbou, and P. Letrémy (2004). Som-based algorithms for qualitative variables. *Neural Networks 17*, 1149–1167.

Cottrell, M. and P. Rousset (1997). The kohonen algorithm: A powerful tool for analysing and representing multidimensional quantitative and qualitative data. In J. Mira, R. Moreno-Dìaz, and J. Cabestany (Eds.), *Biological and Artificial Computation: From Neuroscience to Technology*, Volume 1240 of *Lecture Notes in Computer Science*, pp. 861–871. Springer Berlin / Heidelberg.

Davies, D. L. and D. W. Bouldin (1979). A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell. 1*, 224–227.

Day, W. (1986). Consensus classification. *Journal of Classification Special issue 3(2)*, 345–351.

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

Diday, E. and J. Simon (1976). Clustering analysis. *digital pattern classification Springer Verlag NJ*, 47–94.

Escofier, B. and J. Pagès (1994). Multiple factor analysis (afmult package). *Computational Statistics and Data Analysis 18*(1), 121 – 140.

Ganti, V., J. Gehrke, and R. Ramakrishnan (1999). Cactus-clustering categorical data using summaries. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, New York, NY, USA, pp. 73–83. ACM.

Green, P., A. Krieger, and C. Schaffer (1993). An empirical test of optimal respondent weighting in conjoint analysis. *Journal of the Academy of Marketing Science 21*, 345–351.

Guha, S., R. Rajeev, and S. Kyuseok (2000). Rock: A robust clustering algorithm for categorical attributes. *Information Systems 25*(5), 345–366.

Halkidi, M., Y. Batistakis, and M. Vazirgiannis (2001). Clustering algorithms and validity measures. In *Proceedings of the 13th International Conference on Scientific and Statistical Database Management*, pp. 3–22. IEEE Computer Society.

Heskes, T. (2001). Self-organizing maps, vector quantization, and mixture modeling. *IEEE Transactions on Neural Networks 12*, 1299–1305.

Hollander, M. (1973). *Nonparametric statistical methods*. John Wiley.

Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values. In *In The First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 21–34.

Huang, Z. and M. Ng (1999). A fuzzy k-modes algorithm for clustering categorical data. *Fuzzy Systems, IEEE Transactions on 7*(4), 446 –452.
English

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification 2*, 193–218.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters 31*(8), 651 – 666.

Jing, L., J. Huang, M. Ng, and H. Rong (2004). A feature weighting approach to building classification models by interactive clustering. In *Modeling Decisions for Artificial Intelligence*,

Volume 3131, pp. 1–20. Springer Berlin / Heidelberg.

Johansson, S., M. Jern, and J. Johansson (2008). Interactive quantification of categorical variables in mixed data sets. In *Proceedings of the 2008 12th International Conference Information Visualisation*, Washington, DC, USA, pp. 3–10. IEEE Computer Society.

Jollois, F.-X. and M. Nadif (2002). Clustering large categorical data. In M.-S. Chen, P. Yu, and B. Liu (Eds.), *Advances in Knowledge Discovery and Data Mining*, Volume 2336 of *Lecture Notes in Computer Science*, pp. 257–263. Springer Berlin / Heidelberg.

Kaski, S. (1997). Data exploration using self-organizing maps. *Computational Statistics and Data Analysis. In Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series 1*(82), 57 pp.

Kirchner, S., M. Derbez, C. Duboudin, P. Elias, J. Lucas, N. Pasquier, and O. Ramalho (2009). indoor air quality in french dwellings. *AIVC contributed Report*, 30 p.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics 43*, 59–69.

Kohonen, T. (1995). *Self-organizing maps*. Berlin: Springer Verlag.

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

Kotsiantis, S. B. and P. E. Pintelas (2004). Recent advances in clustering: A brief survey. *WSEAS Transactions on Information Science and Applications 1*, 73–81.

Lebart, L., M. Piron, and A. Morineau (2006). *Statistique exploratoire multidimensionnelle Visualisation et inférence en fouille de données*, Volume 4. Dunod.

Lebbah, M. (2003). *Carte topologique pour données qualitatives : application à la reconnaissance automatique de la densité du trafic routier*. Thèse de doctorat, Université de Versailles Saint Quentin en Yvelines.

Lebbah, M., A. Chazotte, F. Badran, and S. Thiria (2005). Mixed topological map. *ESANN 17*, 357–362.

Lebbah, M., S. Thiria, and F. Badran (2000). Topological map for binary data. *ESANN N*, 26–28.

Luttrell, S. P. (1994). A bayesian analysis of self-organizing maps. *Neural Comput. 6*(5), 767–794.

Milligan, G. and M. Cooper (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika 50*, 159–179.

Rogovschi, N., M. Lebbah, and N. Grozavu (2011). Pondération et classification simultanée de données binaires et continues. *EGC 11*, 65–70.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20*(0), 53 – 65.

Tenenhaus, M. and F. Young (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika 50*, 91–119.

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

Verbeek, J., N. Vlassis, and B. Krase (2005). Self-organizing mixture models. *Neurocomputing 63*(0), 99 – 123.

Vichi, M. (1998). Principal classifications analysis: a method for generating consensus dendrograms and its application to three-way data. *Computational Statistics 27(3)*, 311–331.

Vichi, M. (1999). One-mode classification of a three-way data matrix. *Journal of Classification 16*, 27–44.

Vichi, M. and H. A. Kiers (2001). Factorial k-means analysis for two-way data. *Computational Statistics and Data Analysis 37*(1), 49–64.

Westerhuis, J. A., T. Kourti, and J. F. MacGregor (1998). Analysis of multiblock and hierarchical pca and pls models. *Journal of Chemometrics 12*(5), 301–321.

Wold, S., N. Kettaneh, and K. Tjessem (1996). Hierarchical multiblock pls and pc models for easier model interpretation and as an alternative to variable selection. *Journal of Chemometrics 10*(5-6), 463–482.

Appendix

Description of variables

	Variables	Description
Dwellings	HSRF	Surface
	NIACE	Year of building
	NSQ31	Number of independent combustion equipment connected to a duct for smoked in housing
	TMMB2	Attendance rate of chipboard furniture in the housing (living rooms)
	TRS2	Rate of soil in plastic housing (living rooms)
	TMMB1	Attendance rate of solid wood furniture in the housing (living rooms)
	TMM1	Rate woodwork in the housing (living rooms)
	TMM3	Rate wood other than millwork and PVC in the housing (living rooms)
Household	AGE	Age of householder
	NBP	Number of people living in housing
	REV	Incomes
	NBPEI10	Number of children under 10 years old
	NBPES10	Number of children over 10 years old
Habits	ACT	Maintains overall housing
	CUI1	During the week, how many times have you cooked food in the house?
	QPPB2	During the week, have you used in your home pesticides insecticides as aerosol (for plants, animals, household)?
	QPD2b	During the week, have you used in your home another type of deodorant (diffuser wick candle, lamp, incense, potpourri, air freshener for vacuum cleaner, toilet block, solid, gel, etc..
	QME1+	During the week, how many times have you clean soil floorclothes?

Table 3 – *Main significant variables*