Hierarchical mixed topological maps

N.Niang[1] & M. Ouattara[2]

[1] Département Imath & CEDRIC, CNAM, Paris, FRANCE
[2] Centre Scientifique et Technique du Bâtiment

ndeye.niang_keita@cnam.fr , mory.ouattara@cstb.fr

Outline

1. Introduction
2. Mixed topological maps
3. Hierarchical Mixed topological maps
4. Case study
5. Conclusion & future work

1. Introduction

Context

- A national representative survey conducted by the OQAI (http://www.air-interieur.org/)

- Original study purposes : investigate links between various pollutants (VOC) and several others variables and to classify dwellings across France on their air quality and find which factors influence this quality

Context

- Data collected on several aspects of the dwellings themselves and households living in such as
  - Type of households (marital status, income,...)
  - Technical characteristics of the dwellings
  - Habits of inhabitants (smoke, ...)
  - Pollutants (Formaldheyde, benzene)
- Blocks of mixed variables

Context

Several blocks of mixed variables

	Block 1		...	Block P	
	Quantitative	Qualitative		Quantitative	Qualitative
1	$X_{1r_1} \dots X_{1r_{p1}}$	$X_{1b_1} \dots X_{1b_{g1}}$		$X_{1r_1} \dots X_{1r_{pp}}$	$X_{1b_1} \dots X_{1b_{qp}}$
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
n	$X_{nr_1} \dots X_{nr_{p1}}$	$X_{1b_1} \dots X_{1b_{q1}}$		$X_{nr_1} \dots X_{nr_{pp}}$	$X_{1b_1} \dots X_{1b_{qp}}$

Context

- Preliminary study purpose (from OQAI):
  - Find a clustering of dwellings specific to each block
  - Integrated analysis to have a global synthesis using all the available information
- Problem : two-level clustering with mixed variables structured in blocks

Clustering with mixed variables

- Clustering on principal components from
  - Mutliple factor analysis (Pagès)
  - Categorical principal component analysis (Tenenhaus M.)

- Reduce and cluster simultaneously
  - Factorial K-Means (Vichi & Kiers)
  - Reduced K-means (De Soete & Carroll)

- **Our proposition: Hierarchical MTM**

2. Mixed topological maps

Kohonen self-organized maps

- Neural network unsupervised learning method
- Achieves both tasks of projection and clustering
- Allows visualization of clusters
- SOM consists of neurons organized on a regular two dimensions grid C called map

Kohonen self-organized maps

- Undirected graph
  - Distance δ(c,r): length of the shortest path on C between cells c and r
  - Neighborhood relation defined by a kernel function based on δ and parameterized by T to control the size of the neighborhood

- The neurons are connected to adjacent neurons by the neighborhood relation and that yields the structure of the map on which similar objects should be close together on the grid

Kohonen self-organized maps

- Training data set $A = \left\{ z_i \in \Re^p , i = 1...n \right\}$
- each cell is associated to referent vector w initialised with random samples from A
- Parameterized Cost function to minimize:

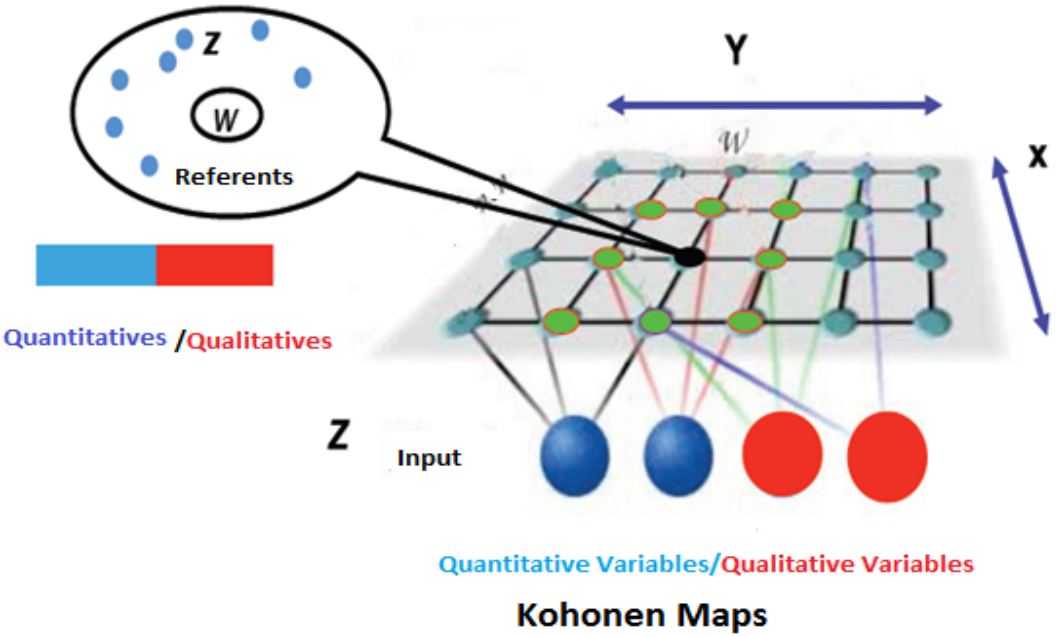
$$J_{SOM}^{T}(\chi, w) = \sum_{z_{i \in A}} \sum_{r \in C} K^{T}(\delta(\chi(z_i), r)) \left\| z_i - w_r \right\|^2$$

- K-means with a weigthed euclidian distance

SOM algorithm: two steps

- **Assigning step** $\chi_T(z) = \arg\min_{r \in C} d_T(z, w_r)$
- Gives a partition of the data $P_c = \{Z \in E / \chi(Z) = c\}$
- **Minimization step** $w_c^T = \dfrac{\sum_{r \in C} K^T(\delta(c,r)) Z_r}{\sum_{r \in C} K^T(\delta(c,r)) n_r}$
- **ONLY for quantitative variables**

Mixed topological maps (Lebbah 2005)



$$P_c = \{Z \in E / \chi(Z) = c\}$$

$$I^T_{MTM}(\chi, w) = \sum_{z_i \in R} \sum_{c \in C} \kappa^T\left(\delta(\chi(z_i), c)\right) * D$$

$$D(z_i, w_c) = \|z_i - w_c\|^2 = \|z_i{}^r - w_c^r\|^2 + \|z_i{}^b - w_c^b\|^2 = \|z_i{}^r - w_c^r\|^2 + 6\mathcal{H}\left(z_i{}^r, w_c^b\right)$$

MTM algorithm

$$w_c^r = \frac{\sum_{z_{i\epsilon E}} \kappa\left(\delta(\chi(z_i), r)\right) z_i{}^r}{\sum_{z_{i\epsilon R}} \kappa\left(\delta(\chi(z_i), r)\right)}$$

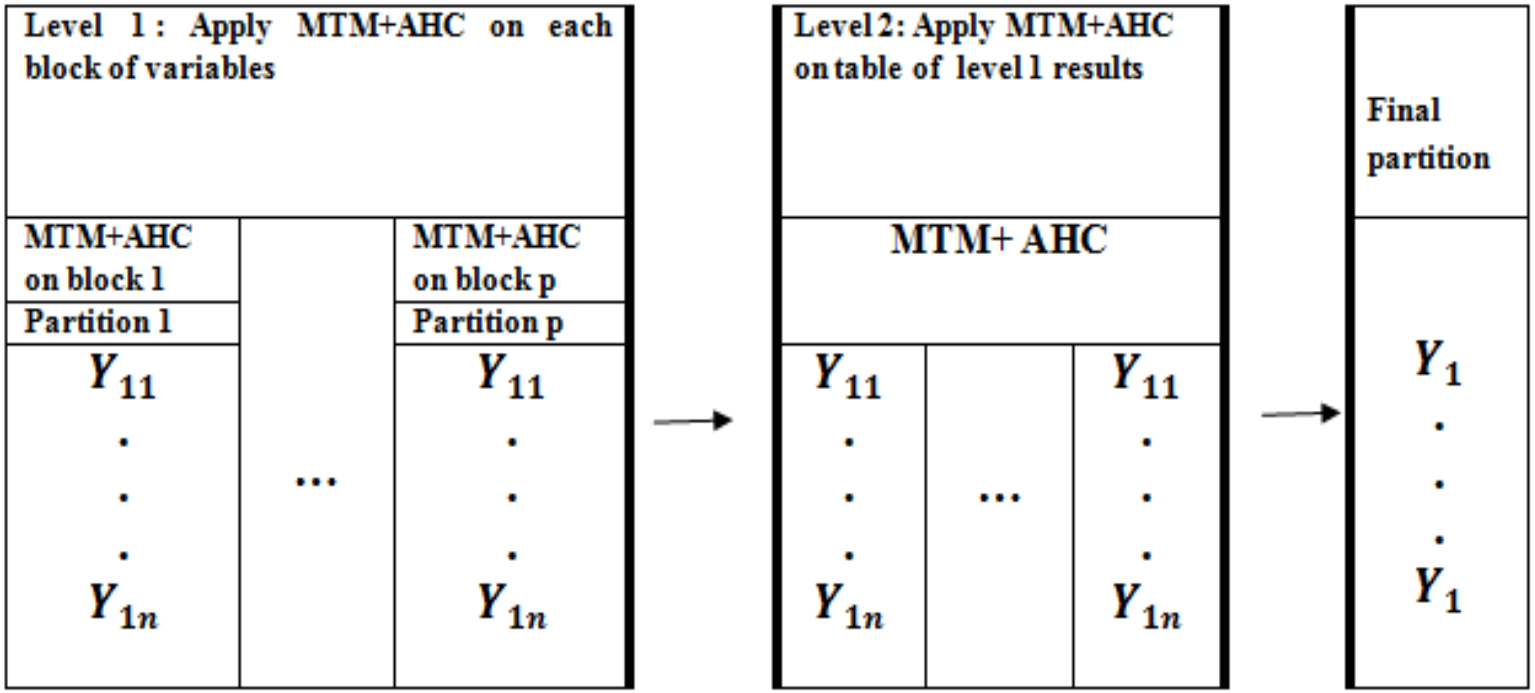
$$w_c^{bk} = \begin{cases} 0 \ si \sum_{z_{i\epsilon A}} \kappa\left(\delta(\chi(z_i), r)\right)\left(1 - z_i{}^{bk}\right) > \sum_{z_{i\epsilon A}} \kappa\left(\delta(\chi(z_i), r)\right) z_i{}^{bk} \\ 1 \ sinon \end{cases}$$

3. Hierarchical mixed topological maps

Our proposition

- Apply MTM to each data set and use AHC
- Apply MTM on the new data set built by horizontal merging of level one results weighted if necessary
- Like Wold's HPCA

Level 2 data set



Level 2 data set

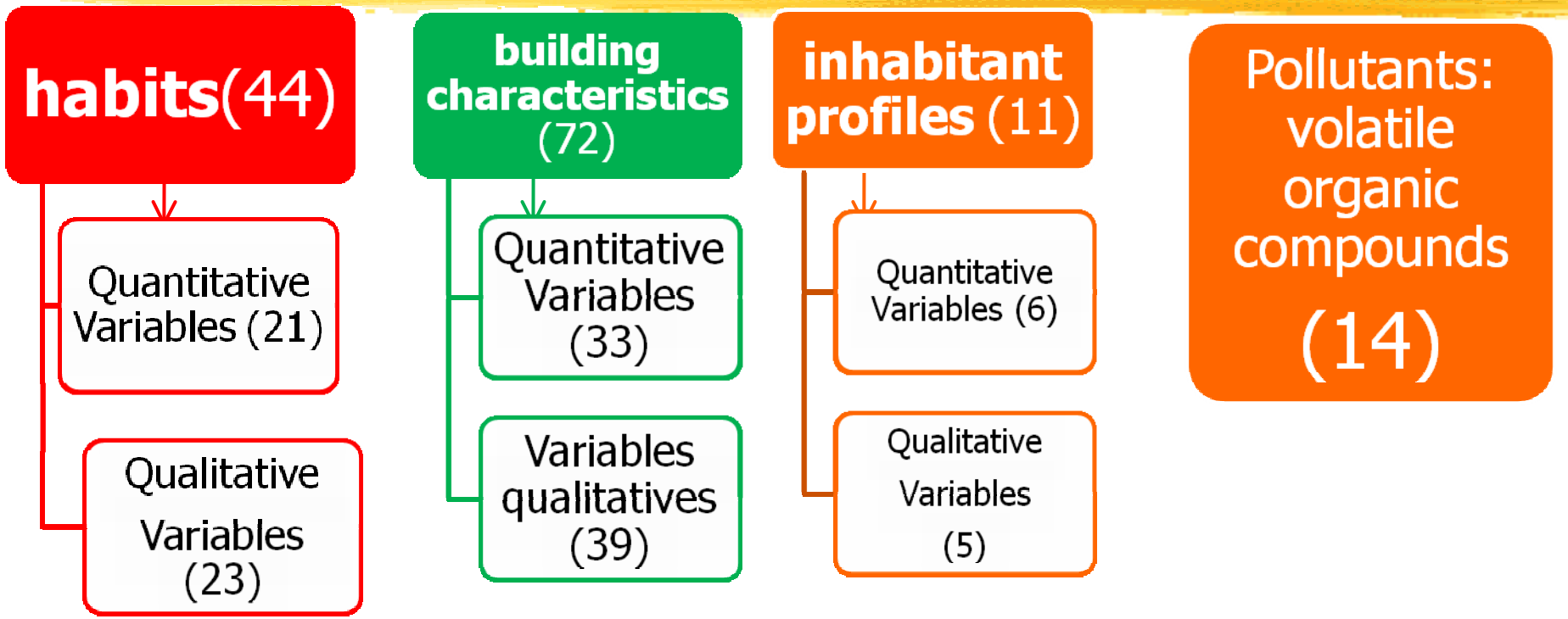
- Or

1	$V_{1\kappa_1} = \left(w_{11} \dots w_{1b_{\kappa_1}}\right)$	...	$V_{1\kappa_b} = \left(w_{1b} \dots w_{1b_{\kappa_b}}\right)$
.	.		.
.	.		.
.	.		.
N	$V_{N\kappa_1} = \left(w_{N1} \dots w_{Nb_{\kappa_N}}\right)$		$V_{N\kappa_b} = \left(w_{N1} \dots w_{Nb_{\kappa_1}}\right)$

4. Case study

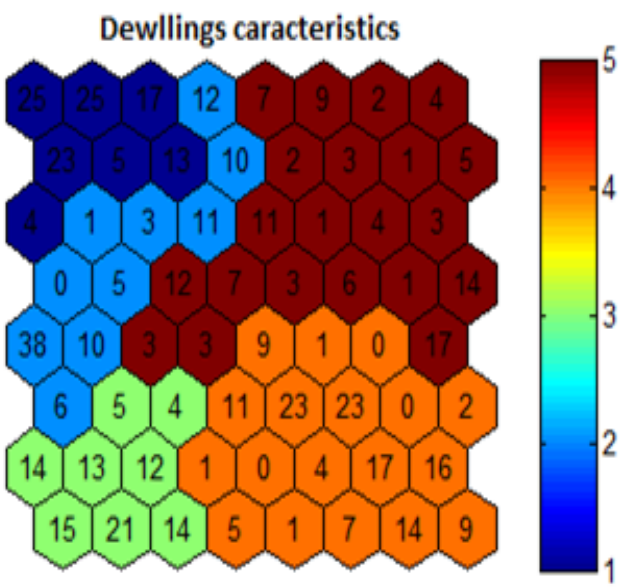
Real data *from OQAI national survey*

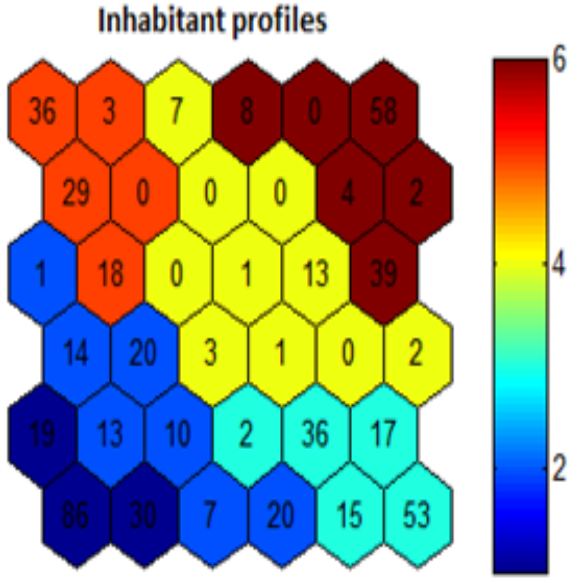
(http://www.air-interieur.org/)

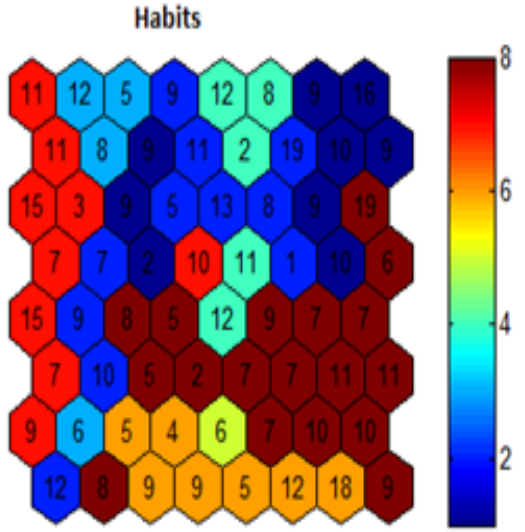


blocks of mixed variables measured on 567 dwellings

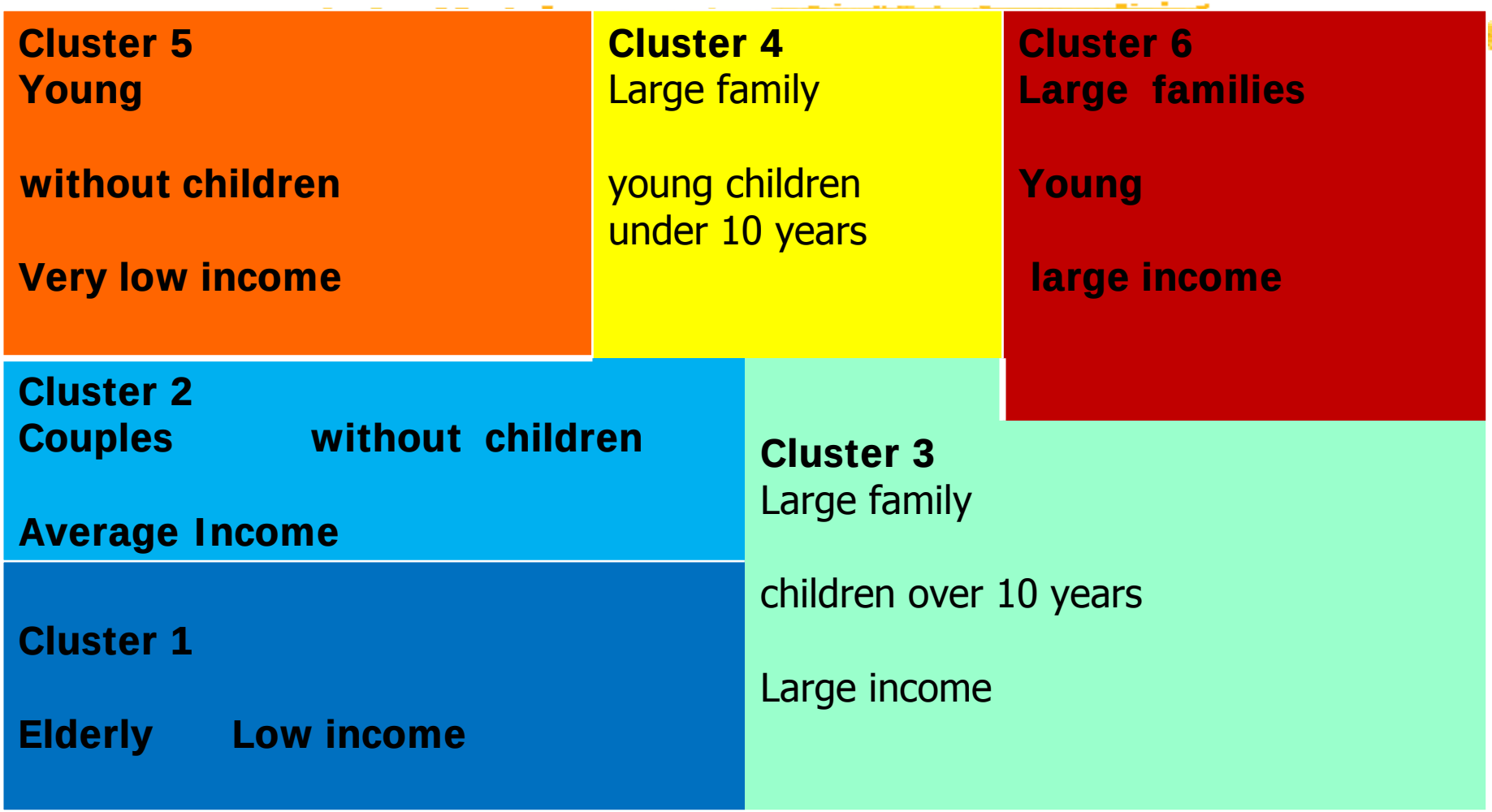
Level 1 : Map of each block of variables after MTM+AHC







Block of inhabitants profiles interpretation



Block of dwellings characteristics interpretation

Cluster 4
Old individual big house all in one
high rate of tiled floor
low rate of carpenter PVC and agglomerated wood
High rate of equipment connected to Smoke conduit

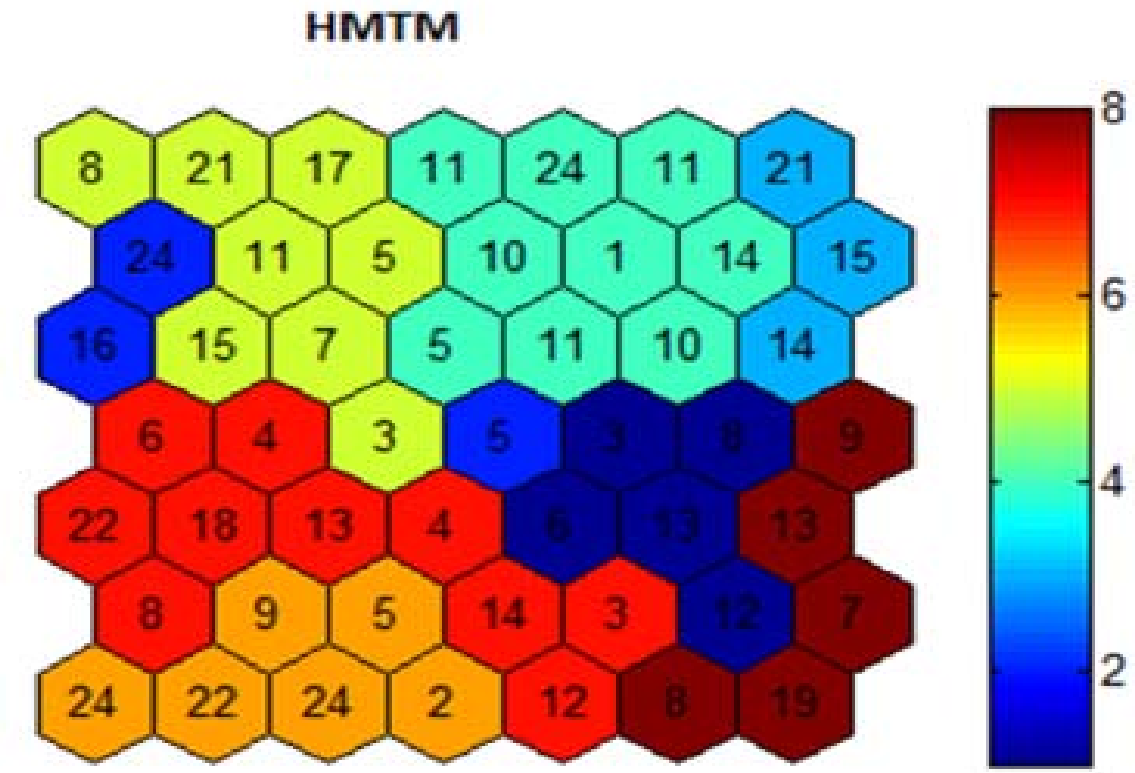
Cluster 5
individual big house all in one
High rate of wood

Cluster 2
Old collectif Small dwellings
Low rate of equipment connected to Smoke conduit
Low rate of tiled floor

Cluster 3
Recents dwellings
high rate of tiled floor
high rate of solid wood furniture
high rate of stratified wooden floors

Cluster 1
Recents collectif Small dwellings
Low rate of equipment connected to Smoke conduit
low rate of solid wood furniture
low rate of tiled floor

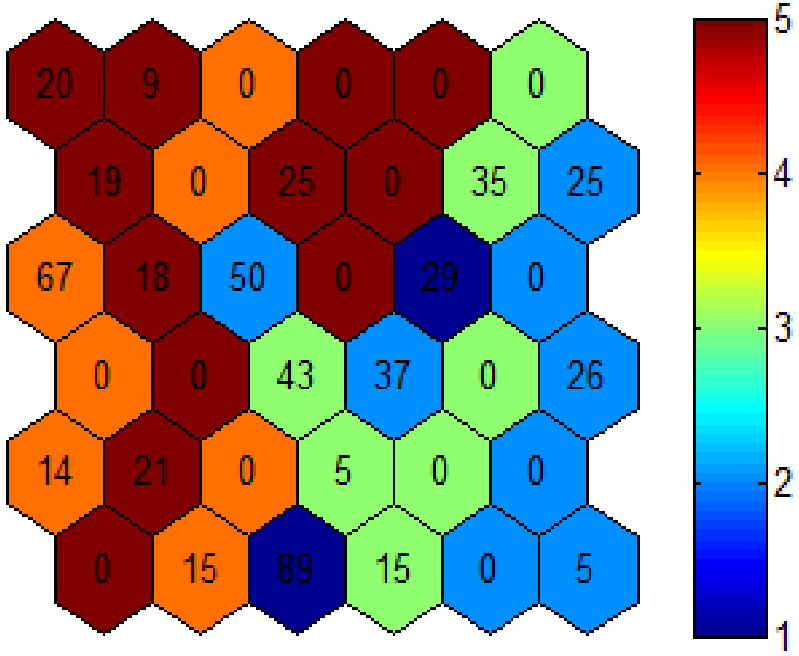
Level 2 : Map of HMTM+AHC on the referents data set



Interpretation with pollutants

	Cl 3 **Big house All in one** He lived alone low incomes Cleaning their home a lot **High concentration of acrolein**	**Cl 4 Big old house All in one** there are couples moderately clean their home **high concentration of toluene**
Cl 6 Recent Collective housing Occupied by young people with two children under 10 years moderately clean their homes low concentration of benzene and Dichlorobenzene		
Cl 7 Recent Collective housing young people with one children over 10 years slightly clean their homes		

Level 2 : Map of HMTM+AHC on the partition data set



Rand index comparison

Rand	Dw Char	Inhabit	Habits	HMTM	MTM
Dw Char	1	0,69	0,70	0,87	0.73
Inhabit		1	0,72	0,74	0.67
Habits			1	0,75	0.67
HMTM				1	0.73
MTM					1

5. Conclusion & future work

Closing Remarks

- We present a two steps method for clustering individuals described by mixed variables structured in homogeneous blocks giving both
  - local synthesis of each block information
  - global summary, consensus of local clustering
- The proposed method applied to indoor air pollution data gives interesting insights for the OQAI

Future work

- Development of MTM: other distance for mixed variables
- Comparative studies with other methods: FKM, RKM, tandem approaches

- Other cluster validity indexes
- Adaptative weights to be found

- Extend the method to indoor air quality of offices
- ……

REFERENCES

- Lebbah M., Chazotte A., Badran F., Thiria S. (2005) Mixed Topological Map, *ESANN 2005, Bruges April*
- Pagès J. (2004) Analyse factorielle de données mixtes, in : *Revue de statistique appliquée,* 52 , 93-111.
- Tenenhaus M., Young F. (1985) An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data, in: *Psychometrika,* 50, 91-119.
- Timmerman M., Ceulemans E., Kiers H., Vichi M. (2010) Factorial and reduced K-means reconsidered, in: *CSDA,* 54, 1858-1871.
- Wold S., Kettaneh N., Tjessem K. (1996) Hierarchical Multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection, in: *Journal of chemometrics,* 10, 463-482.