Hierarchical mixed topological maps

N. Niang\textsuperscript{1} & M. Ouattara\textsuperscript{2}

\textsuperscript{1}Département Imath & CEDRIC, CNAM, Paris, FRANCE
\textsuperscript{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
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 (Formaldehyde, benzene)

Blocks of mixed variables
**Context**

Several blocks of mixed variables

<table>
<thead>
<tr>
<th>Block 1</th>
<th>Block P</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Quantitative</strong></td>
<td><strong>Quantitative</strong></td>
</tr>
<tr>
<td>$X_{1r_1} \ldots X_{1r_{p_1}}$</td>
<td>$X_{1r_1} \ldots X_{1r_{p_p}}$</td>
</tr>
<tr>
<td>$X_{1b_1} \ldots X_{1b_{q_1}}$</td>
<td>$X_{1b_1} \ldots X_{1b_{q_p}}$</td>
</tr>
<tr>
<td><strong>Qualitative</strong></td>
<td><strong>Qualitative</strong></td>
</tr>
<tr>
<td>$X_{n_{r_1}} \ldots X_{n_{r_{p_1}}}$</td>
<td>$X_{n_{r_1}} \ldots X_{n_{r_{p_p}}}$</td>
</tr>
<tr>
<td>$X_{1b_1} \ldots X_{1b_{q_1}}$</td>
<td>$X_{1b_1} \ldots X_{1b_{q_p}}$</td>
</tr>
</tbody>
</table>
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
  - Multiple 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 called map
Kohonen self-organized maps

- Undirected graph
  - Distance $\delta(c,r)$: length of the shortest path on $C$ between cells $c$ and $r$
  - Neighborhood relation defined by a kernel function based on $\delta$ 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 \( A = \{z_i \in \mathbb{R}^p, i=1...n\} \)
- each cell is associated to referent vector \( w \) initialised with random samples from \( A \)
- Parameterized Cost function to minimize:

\[
J^T_{SOM}(\chi, w) = \sum_{z_i \in A} \sum_{r \in C} K^T(\delta(\chi(z_i), r)) \|z_i - w_r\|^2
\]

- K-means with a weightheed 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 $w_c^T = \sum_{r \in C} \frac{K^T(\delta(c,r))Z_r}{\sum_{r \in C} K^T(\delta(c,r))n_r}$

ONLY for quantitative variables
Mixed topological maps (Lebbah 2005)

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

\[ I_{MTM}(\chi, w) = \sum \sum \kappa^T(\delta(\chi(z_i), c)) * D \]

\[ D(z_i, w_c) = \| z_i - w_c \|^2 = \| z_i^r - w_c^r \|^2 + \| z_i^b - w_c^b \|^2 = \| z_i^r - w_c^r \|^2 + 6H(z_i^r, w_c^b) \]
MTM algorithm

\[ w^r_c = \frac{\sum_{z_i \in E} k(\delta(\chi(z_i), r)) z_i^r}{\sum_{z_i \in F} k(\delta(\chi(z_i), r))} \]

\[ w^{bk}_c = \begin{cases} 0 & \text{si} \sum_{z_i \in A} k(\delta(\chi(z_i), r)) (1 - z_i^{bk}) > \sum_{z_i \in A} k(\delta(\chi(z_i), r)) z_i^{bk} \\ 1 & \text{sinon} \end{cases} \]
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 1: Apply MTM+AHC on each block of variables

<table>
<thead>
<tr>
<th>MTM+AHC on block 1</th>
<th>MTM+AHC on block p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partition 1</td>
<td>Partition p</td>
</tr>
<tr>
<td>( Y_{11} )</td>
<td>( Y_{11} )</td>
</tr>
<tr>
<td>( \cdot )</td>
<td>( \cdot )</td>
</tr>
<tr>
<td>( \cdot )</td>
<td>( \cdot )</td>
</tr>
<tr>
<td>( Y_{1n} )</td>
<td>( Y_{1n} )</td>
</tr>
</tbody>
</table>

Level 2: Apply MTM+AHC on table of level 1 results

<table>
<thead>
<tr>
<th>MTM+ AHC</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y_{11} )</td>
</tr>
<tr>
<td>( \cdot )</td>
</tr>
<tr>
<td>( \cdot )</td>
</tr>
<tr>
<td>( Y_{1n} )</td>
</tr>
<tr>
<td>( \cdot )</td>
</tr>
</tbody>
</table>

Final partition

| \( Y_1 \) |
| \( \cdot \) |
| \( \cdot \) |
| \( Y_{1n} \) |

HCSDA-11, October 27th-29th 2011
Beijing, China

Development
Level 2 data set

Or

\[
V_{1\times 1} = (w_{11} \cdots w_{1b_{1\times 1}}) \\
V_{1\times b} = (w_{1b_{1\times b}}) \\
V_{N\times 1} = (w_{N1} \cdots w_{Nb_{N\times 1}}) \\
V_{N\times b} = (w_{N1} \cdots w_{Nb_{N\times b}})
\]
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

Cluster 5
Young
without children
Very low income

Cluster 4
Large family
young children under 10 years

Cluster 6
Large families
Young
large income

Cluster 2
Couples without children
Average Income

Cluster 3
Large family
children over 10 years
Large income

Cluster 1
Elderly Low income
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

<table>
<thead>
<tr>
<th>Cl 3</th>
<th>Big house All in one</th>
</tr>
</thead>
<tbody>
<tr>
<td>He lived alone</td>
<td></td>
</tr>
<tr>
<td>low incomes</td>
<td></td>
</tr>
<tr>
<td>Cleaning their home a lot</td>
<td></td>
</tr>
<tr>
<td>High concentration of acrolein</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cl 4</th>
<th>Big old house All in one</th>
</tr>
</thead>
<tbody>
<tr>
<td>there are couples</td>
<td></td>
</tr>
<tr>
<td>moderately clean their home</td>
<td></td>
</tr>
<tr>
<td>high concentration of toluene</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cl 6</th>
<th>Recent Collective housing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occupied by young people with two children under 10 years</td>
<td></td>
</tr>
<tr>
<td>moderately clean their homes</td>
<td></td>
</tr>
<tr>
<td>low concentration of benzene and Dichlorobenzene</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cl 7</th>
<th>Recent Collective housing</th>
</tr>
</thead>
<tbody>
<tr>
<td>young people with one children over 10 years</td>
<td></td>
</tr>
<tr>
<td>slightly clean their homes</td>
<td></td>
</tr>
</tbody>
</table>
Level 2: Map of HMTM+AHC on the partition data set
## Rand index comparison

<table>
<thead>
<tr>
<th>Rand</th>
<th>Dw Char</th>
<th>Inhabit</th>
<th>Habits</th>
<th>HMTM</th>
<th>MTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dw Char</td>
<td>1</td>
<td>0.69</td>
<td>0.70</td>
<td>0.87</td>
<td>0.73</td>
</tr>
<tr>
<td>Inhabit</td>
<td>1</td>
<td>0.72</td>
<td>0.74</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>Habits</td>
<td>1</td>
<td></td>
<td>0.75</td>
<td></td>
<td>0.67</td>
</tr>
<tr>
<td>HMTM</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>0.73</td>
</tr>
<tr>
<td>MTM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
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