

## Mixed Topological Map

Mustapha. LEBBAH<sup>a,b</sup>, Aymeric. CHAZOTTES<sup>b</sup>  
Fouad. BADRAN<sup>a</sup>, Sylvie. THIRIA<sup>b</sup>

a- CEDRIC, Conservatoire National des Arts et Métiers,  
292 rue Saint Martin, 75003 Paris, France

b- Laboratoire LODYC, Université Paris 6, Tour 45-5<sup>e</sup> étage,  
4 place Jussieu 75252 Paris cedex 05 France

**Abstract.** We propose a new algorithm which is based on a topological map model and dedicated to mixed data, with numerical and binary components. The algorithm computes directly the referent vectors, as mixed data vectors sharing the same interpretation with the observations. The method is validated on a real data related to the ocean colour domain.

### 1 Introduction

The topological map proposed by Kohonen [8] uses a self-organization algorithm (SOM) which provides quantification and clustering of the observation space. More recently, new models of topological maps dedicated to specific data were proposed in [2, 4, 5, 6, 7]. Some of these models are based on a probabilistic formalism and a learning procedure to maximize the likelihood function of the data set, the others are quantization methods. In this paper we proposed a topological self organizing map for analysing mixed (numerical and binary) data. It is a quantization model which provides a consistent set of prototypes whose particularity is to be interpreted (prototypes and data belong to the same space and have a meaningful interpretation). In section 2, we present the model and the iterative algorithm. In section 2.2, we validate the model on a real problem for which the observations are mixed data (numerical and binary).

### 2 Mixed Topological Map (MTM)

Let  $A = \{\mathbf{z}_i, i = 1..I\}$  be the learning data set where each observation  $\mathbf{z}_i = (z_i^1, z_i^2, \dots, z_i^d)$  is made of two parts: numerical part  $\mathbf{z}_i^r = (z_i^{1r}, z_i^{2r}, \dots, z_i^n)$  ( $\mathbf{z}_i^r \in \mathcal{R}^n$ ) and binary part  $\mathbf{z}_i^b = (z_i^{1b}, z_i^{2b}, \dots, z_i^m)$  ( $\mathbf{z}_i^b \in \beta^m = \{0, 1\}^m$ ). With these notations a particular observation  $\mathbf{z}_i = (\mathbf{z}_i^r, \mathbf{z}_i^b)$  is a mixed vector (numerical and binary) of dimension  $d = n + m$ . As for the classical topological maps, we assume that we have regular grid  $\mathcal{C}$ , on which we define distance  $\delta(c, r)$ , as the length of the unique shortest path on the grid (as undirected graph) between the cells  $c$  and  $r$ . The neighborhood system can be defined by a monotonically decreasing kernel function  $\mathcal{K}$  ( $\mathcal{K} \geq 0$  and  $\lim_{|x| \rightarrow \infty} \mathcal{K}(x) = 0$ ). We define the mutual influence of two cells  $c$  and  $r$  by  $\mathcal{K}(\delta(c, r))$ . For each cell  $c$  of the grid, we associate a referent vector  $\mathbf{w}_c = (\mathbf{w}_c^r, \mathbf{w}_c^b)$  of dimension  $d$ , where  $\mathbf{w}_c^r \in \mathcal{R}^n$

and  $\mathbf{w}_c^b \in \beta^m$ . We denote by  $\mathcal{W}$  the set of the referent vectors, by  $\mathcal{W}^r$  the set of the numerical part and by  $\mathcal{W}^b$  the binary part of the referent vectors. In the following section we present a new model of topological map dedicated to mixed data. The associated learning algorithm is derived from the batch version of the Kohonen algorithm dedicated to numerical data [8] and the BinBatch algorithm which is dedicated to binary data [7]. In this algorithm, the similarity measure and the estimation of the referent vectors are specific in each data set: it is the Euclidian distance with the mean vector in the numerical case and the Hamming distance with the median center in the binary case.

## 2.1 Minimization of the cost function

As the classical topological maps we propose to minimize the following cost function [1, 8].

$$\mathcal{E}(\phi, \mathcal{W}) = \sum_{\mathbf{z}_i \in App} \sum_{r \in \mathcal{C}} \mathcal{K}(\delta(\phi(\mathbf{z}_i), r)) \|\mathbf{z}_i - \mathbf{w}_r\|^2 \quad (1)$$

Where  $\phi$  assigns each observation  $\mathbf{z}$  to a single cell in  $\mathcal{C}$ . In this expression  $\|\mathbf{z} - \mathbf{w}_r\|^2$  is square of the Euclidian distance. Since for binary vectors the Euclidian distance is no more than the Hamming distance  $\mathcal{H}$ , then the Euclidian distance can be rewritten by:  $\|\mathbf{z} - \mathbf{w}_r\|^2 = \|\mathbf{z}^r - \mathbf{w}_r^r\|^2 + \mathcal{H}(\mathbf{z}^b, \mathbf{w}_r^b)$ . Using this expression, the cost function is

$$\begin{aligned} \mathcal{E}(\phi, \mathcal{W}) &= \sum_{\mathbf{z}_i \in App} \sum_{r \in \mathcal{C}} \mathcal{K}(\delta(\phi(\mathbf{z}_i), r)) [\mathcal{D}_{euc}(\mathbf{z}_i^r, \mathbf{w}_r^r) \\ &\quad + \sum_{\mathbf{z}_i \in App} \sum_{r \in \mathcal{C}} \mathcal{K}(\delta(\phi(\mathbf{z}_i), r)) \mathcal{H}(\mathbf{z}_i^b, \mathbf{w}_r^b)] \quad (2) \\ &= \mathcal{E}_{som}(\phi, \mathcal{W}^r) + \mathcal{E}_{bin}(\phi, \mathcal{W}^b) \end{aligned}$$

Where

$$\mathcal{E}_{som}(\phi, \mathcal{W}) = \sum_{\mathbf{z}_i \in App} \sum_{r \in \mathcal{C}} \mathcal{K}(\delta(\phi(\mathbf{z}_i), r)) \|\mathbf{z}_i^r - \mathbf{w}_r^r\|^2 \quad (3)$$

is the classical cost function used by the Kohonen Batch algorithm [8], and

$$\mathcal{E}_{bin}(\phi, \mathcal{W}) = \sum_{\mathbf{z}_i \in App} \sum_{r \in \mathcal{C}} \mathcal{K}(\delta(\phi(\mathbf{z}_i), r)) \mathcal{H}(\mathbf{z}_i^b, \mathbf{w}_r^b) \quad (4)$$

is the cost function used in BinBatch algorithm [7]. The minimization of the cost function (1), is made using an iterative process with two steps:

- **Assignment step:** assuming that  $\mathcal{W}$  is fixed, we have to minimize  $\mathcal{E}(\phi, \mathcal{W})$  with respect to  $\phi$ . This leads to use the following assignment function:  $\forall \mathbf{z}, \phi(\mathbf{z}) = \arg \min_c (\|\mathbf{z}^r - \mathbf{w}_c^r\|^2 + \mathcal{H}(\mathbf{z}^b, \mathbf{w}_c^b))$ .
- **Optimization step:** assuming that  $\phi$  is fixed, this step minimizes  $\mathcal{E}(\phi, \mathcal{W})$  with respect to  $\mathcal{W}$  in the space  $R^n \times \beta^m$ . The minimization of the cost

function (1) leads to minimize the function  $\mathcal{E}_{som}(\phi, \mathcal{W})$  (3) in  $R^n$  and  $\mathcal{E}_{bin}(\phi, \mathcal{W})$  (4) in  $\beta^m$ . It is easy to see that this two minimizations allow to define:

- the numerical part  $\mathbf{w}_c^r$  of the referent vector  $\mathbf{w}_c$  as the mean vector as:

$$\mathbf{w}_c = \frac{\sum_{\mathbf{z}_i \in \mathcal{A}} \mathcal{K}(\delta(c, \phi(\mathbf{z}_i))) \mathbf{z}_i^r}{\sum_{\mathbf{z}_i \in \mathcal{A}} \mathcal{K}(\delta(c, \phi(\mathbf{z}_i)))},$$

- the binary part  $\mathbf{w}_c^b$  of the referent vector  $\mathbf{w}_c$  as the median center of the binary part of the observations  $\mathbf{z}_i \in \mathcal{A}$  weighted by  $\mathcal{K}(\delta(c, \phi(\mathbf{z}_i)))$ . Each component  $\mathbf{w}_c^b = (w_c^{b1}, \dots, w_c^{bk}, \dots, w_c^m)$  is then computed as follows:  $w_c^{bk} =$

$$\begin{cases} 0 & \text{if } \left[ \sum_{\mathbf{z}_i \in \mathcal{A}} \mathcal{K}(\delta(c, \phi(\mathbf{z}_i))) (1 - \mathbf{z}_i^{bk}) \right] \geq \\ & \left[ \sum_{\mathbf{z}_i \in \mathcal{A}} \mathcal{K}(\delta(c, \phi(\mathbf{z}_i))) \mathbf{z}_i^{bk} \right] \\ 1 & \text{otherwise} \end{cases},$$

The minimization of  $\mathcal{E}(\phi, \mathcal{W})$  is run by iteratively performing the two steps. At the end  $\mathbf{w}_c$ , which shares the same code with the observations can be decoded in the same way, allowing a symbolic interpretation of binary part of referent vectors. The nature of the topological model reached at the end of the algorithm, the quality of the clustering and those of the topological order induced by the graph greatly depend on the neighborhood function  $\mathcal{K}$ . In practice, as for traditional topological map we use smooth function to control the size of the neighborhood as  $\mathcal{K}^T(\delta(c, r)) = \exp\left(\frac{-\delta(c, r)}{T}\right)$ . Using this kernel function,  $T$  becomes a parameter of the model. As in the Kohonen algorithm [8], we repeat the preceding iterations by decreasing  $T$  from an initial value  $T_{max}$  to a final value  $T_{min}$ .

## 2.2 Experiments on Ocean colour

In the following we present some experiments to illustrate the convergence of the Mixed Topological Map (MTM) algorithm dedicated to mixed data and we discuss the interesting aspect of the approach. We deal with a real problem concerning the ocean colour domain. We first present the data. We then consider the result of MTM learning. Finally we give a comparison to SOM learning.

### 2.2.1 The data set

The spectral values of light absorption by algal populations are strongly dependent on the pigment composition of the phytoplankton assemblage and therefore contain information on their taxonomic composition. We believe that using the MTM algorithm with adequate qualitative criteria will help to strengthen these differences. Data are made of samples collected during ten cruises, in various seasons and various areas of the world ocean [3]. A sample  $\mathbf{z}_i = (z_i^1, z_i^2, \dots, z_i^{40})$  is compounded both with numerical components  $\mathbf{z}_i^r = (z_i^{1r}, z_i^{2r}, \dots, z_i^{31r})$  and with binary categorical attributes  $\mathbf{z}_i^b = (z_i^{1b}, z_i^{2b}, \dots, z_i^{9b})$ :

- $\mathbf{z}_i^r$  is a 31 component vector corresponding to the coding of a sample absorption spectrum. The first component  $z_i^{1r}$  stands for the amplitude of the spectrum, the 30 others ( $z_i^{2r}, \dots, z_i^{31r}$ ) for the slopes of the spectrum. All components of  $\mathbf{z}_i^r$  are normalised between 0 and 1.

- $\mathbf{z}_i^b$  is a nine-dimensional vector which represents 3 categorical variables: the Size (Small, Medium, Large), the Water Type (Poor, Medium, Rich) and the Chlorophyll B Proportion (Weak, Normal, Strong). A  $\mathbf{z}_i$  stands for a sample of ocean water. In this sample, numerous phytoplankton cells with various sizes can be represented. For an ocean water sample, we establish the Size categorical variable by choosing the modality represented by more than 50%. Thus a sample can be represented by (0,0,0) meaning that the ratio is never greater than 50%.

The database (2163 data points) is finally split at random in two sets: the learning set *App* (1728 data, 80% of the total data set) and the test set *Test* (435 data, 20% of the total data set).

### 2.2.2 Application of the MTM algorithm

In this part we illustrate the convergence of the MTM algorithm on a  $10 \times 10$  map. The figure 1 presents the categorical components and some numerical components for the referent vectors. Moreover, some biological characteristics can be retrieved by the MTM map.

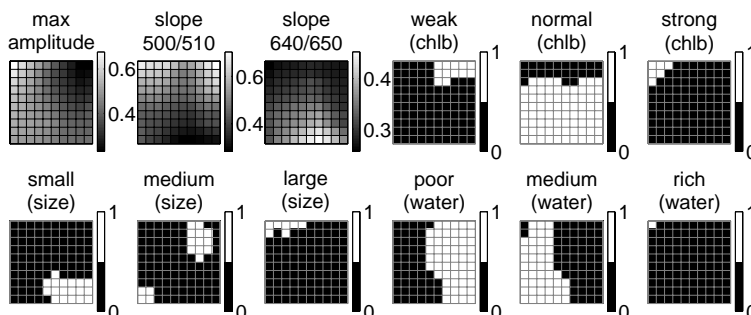


Fig. 1: These 12 pictures represent the different components of the  $10 \times 10$  prototypes. The three pictures in the upper left side of the figure stands for the numerical components (max amplitude and 2 spectrum slopes). The 9 other pictures represent all the binary components (white stands for “1” and black for “0”).

The amplitude of the spectrum covaries with the water type. This can be observed when looking at the max amplitude map (top left corner) and the water type categorical variable (right of the lower line). The map presents a well organized topological order. The use of categorical variables allows to retrieve an order linked with these variables : for exemple “Large Size” data are located in the same corner as a compact cluster.

In table 1, we compare, for the “Size” categorical variable, the labels of the MTM map referent vectors to the result of the majority vote realized on each neuron subset of observations. Clearly the results are coherent and stable for the generalisation.

Adding binary components to numerical data moves away data belonging to different classes even though their numerical components are close.

Small (Learn)			Medium (Learn)			Large (Learn)		
	$\neg$ Small	Small		$\neg$ Med.	Med.		$\neg$ Large	Large
$C_{\neg$ Small}	78.8%	3.1%	$C_{\neg$ Med.	73.0%	1.4%	$C_{\neg$ Large}	78.4%	7.9%
$C_{Small}$	0.0%	18.1%	$C_{Med.}$	0.0%	25.6%	$C_{Large}$	0.7%	13.0%

Small (Test)			Medium (Test)			Large (Test)		
	$\neg$ Small	Small		$\neg$ Med.	Med.		$\neg$ Large	Large
$C_{\neg$ Small}	81.4%	2.3%	$C_{\neg$ Med.	68.7%	3.2%	$C_{\neg$ Large}	79.5%	7.8%
$C_{Small}$	0.0%	16.3%	$C_{Med.}$	0.9%	27.1%	$C_{Large}$	0.5%	12.2%

Table 1: Confusion matrix for the **Size** categorical variable (Learn and Test).  $C_*$  and  $C_{\neg*}$  stands for the MTM classification where  $\neg*$  signifies that the ratio of  $*$  ( $*$  = small, medium or large) particles is less than 50%

### 2.2.3 Coherence between numeric and binary components

In this section, we test the robustness of the MTM map to the missing data. As the knowledge of the 3 categorical data require a lot of biological analysis we project the data according to their numerical components only (the binary part is not used). For each observation of the learning set we assign the observation to its “nearest numerical neuron”. Then the cells are labelled according to the categorical part of their captured spectra (each cell computes the median center of its observations). Figure 2 shows, for the size variable, the original MTM map (component of the referent vectors, left) and the results of the labelling process for the learning set (middle) and the validation test (right).

When we project the  $\mathbf{z}_i^r$  we loose the information contained in the binary data. Figure 2 allows to conclude that the binary components are coherent with the numerical components: the information of the categorical data can be retrieved using only the numerical part of the observation. Clearly, patterns of the six figures on the right correspond to the three figures on the left ones. Nevertheless, we notice that some differences are due to the information not included in the numerical vector.

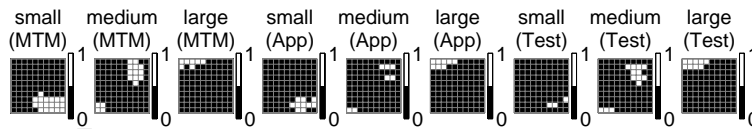


Fig. 2: MTM map referents with Learn and Test “Size” label.

By comparing the performances of the confusion matrix in the Tables 2 and 1, we can estimate the accuracy of the MTM map.

Small (Learn)			Medium (Learn)			Large (Learn)		
	$\neg$ Small	Small		$\neg$ Med.	Med.		$\neg$ Large	Large
$C_{\neg$ Small}	72.9%	10.5%	$C_{\neg$ Med.	71.5%	20.8%	$C_{\neg$ Large}	77.4%	8.1%
$C_{Small}$	6.0%	10.8%	$C_{Med.}$	1.4%	6.3%	$C_{Large}$	1.6%	12.8%

Small (Test)			Medium (Test)			Large (Test)		
	$\neg$ Small	Small		$\neg$ Med.	Med.		$\neg$ Large	Large
$C_{\neg$ Small}	77.5%	10.6%	$C_{\neg$ Med.	66.7%	18.9%	$C_{\neg$ Large}	78.2%	7.6%
$C_{Small}$	3.9%	8.0%	$C_{Med.}$	3.0%	11.5%	$C_{Large}$	1.8%	12.4%

Table 2: Confusion matrix for the **Size** categorical variable (Learn and Test projected on the numerical components)

#### 2.2.4 Contribution against the som algorithm

In this last experiment we trained a classical Self Organizing Map (SOM). We projected the learning and the test sets and applied a majority vote as in the section 2.2.3. The result of their vote for the learning (left) and the test set (right) is presented in figure 3. Clearly, the SOM map exhibit less structure, particularly for the medium size which has not even a “winner” for the test.

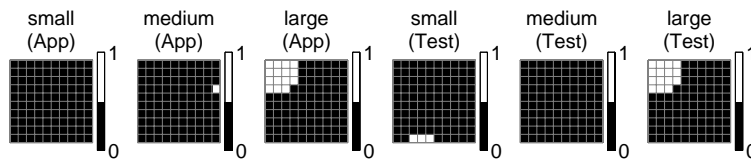


Fig. 3: SOM map Learn and Test “Size” label.

### 3 Conclusion

We present in this paper a new algorithm (MTM) which is based on a topological map model and dedicated to mixed data, with numerical and binary components. This algorithm uses simultaneously the Euclidian and the Haming distance on the two data components to compute referent vectors with mixed components, sharing the same interpretation of the observations. We illustrate the behaviour of the MTM algorithm on a real data concerning the ocean colour domain. The analysis of the result shows the possibility of the resulting Map : in topology preserving order, in the reconstitution of binary components from the numerical ones and in clustering.

**Acknowledgements:** This work was supported by the European Commission Research Directorate-General (NAOC, contract no. EVG1-CT-2000-00034). We wish to thank the Dr. A. Bricaud from LOV (Laboratoire d’Océanographie de Villefranche) who provided us the large biogeochemical and optical data set.

### References

- [1] Anouar F. Badran, F. Thiria, S. 1998. Probabilistic self-organizing map and radial basis function networks. *Neurocomputing*
- [2] Bishop, C. M., Svensen, M., and Williams, C. K. I. (1998): GTM: The Generative Topographic Mapping. *Neural Computation*, 10(1), 215-234
- [3] Bricaud A., H. Claustre, J. Ras, K. Oubelkheir: (2004) Natural variability of phytoplanktonic absorption in oceanic waters: influence of the size structure of algal populations, *J. Geophys. Res.*, Vol. 109, No. C11, C11010
- [4] Kaban, A and Girolami, M. (2001): A Combined Latent Class and Trait Model for the Analysis and Visualisation of Discrete Data. *I.E.E.E Transactions on Pattern Analysis and Machine Intelligence*.23(8), pp859 -872
- [5] Ibbou, S. Cottrell. M 1995 *Multiple correspondense Analysis crosstabulation matrix: using the Kohonen algorithm*. In Verleysen, M. Editor proc of ESANN'95, pages 27-32. Dfacto Bruxelles
- [6] Lebbah. M, Thiria. S, Badran. ESANN, Visualization and clustering with Categorical Topological Map, ESANN 2004, Bruges, April 26-27-28, 2004, Proceedings
- [7] Lebbah. M, Thiria. S, Badran. ESANN, Topological Map for Binary Data, ESANN 2000, Bruges, April 26-27-28, 2000, Proceedings
- [8] Kohonen, T. 1994. *Self-Organizing Map*. Springer, Berlin