Applied Stochastic Model and Data Analysis. Brest (France) May 17-20, 2005.

Visualization of textual data : unfolding the Kohonen maps.

Ludovic Lebart, CNRS and Télécom Paris. lebart@enst.fr

Abstract: The Kohonen self organizing maps (SOM) can be viewed as a visualisation tool that performs a sort of compromise between a high-dimensional set of clusters and the 2-dimensional plane generated by some principal axes techniques. The paper proposes, through Contiguity Analysis, a set of linear projectors providing a representation as close as possible to a SOM map. As expected, owing to the non-linear character of the representation, such projectors will only concern local parts of the SOM maps. In so doing, we can obtain an idea of the variability of points representing words via a standard partial bootstrap procedure.

1. Introduction

For many users of visualisation tools, the Kohonen self organising maps (SOM) outperform both usual clustering techniques and principal axes techniques (principal components analysis, correspondence analysis, etc.). On the one hand, the displays of identifiers of words (or text units) within rectangular or octagonal cells allow for clear and legible printings. On the other hand, the SOM grid, basically non-linear, can be viewed as a compromise between a high-dimensional set of clusters and the 2-dimensional plane generated by any pairs of principal axes. One can regret however the absence of assessment procedures and of valid statistical inference as well. The paper proposes, through Contiguity Analysis (briefly reminded in section 2), a set of linear projectors providing a representation as close as possible to a SOM map (section 3 and 4). An example of application is given in section 5. As expected, owing to the non-linear character of the representation, such projectors will only concern local parts of the SOM maps. In so doing, we can obtain an idea of the variability of points representing words via a standard partial bootstrap procedure. We can then provide the SOM maps with the projection of confidence areas (e.g. ellipses) around the location of words (section 6).

2. Brief reminder about contiguity analysis

Let us consider a set of multivariate observations, (n observations described by p variables, leading to a (n,p) matrix **X**), having an *a priori* graph structure. The n observations are also the n vertices of a symmetric graph G, whose associated matrix is **M** ($m_{ii'} = 1$ if vertices i and i' are joined by an edge, $m_{ii'} = 0$ otherwise). We denote by **N** the (*n*,*n*) diagonal matrix having the degree of each vertex i as diagonal element n_i (n_i stands here for $n_{ii'}$). *y* is the vector whose i-th components is y_i. Note that: $n_i = \Sigma_{i'} m_{ii'}$. **U** designates the square matrix such that $u_{ij} = 1$ for all i and j.

2.1. Local variance $v^c(y)$ of a variable y

y being a random variable taking values on each vertex i of a symmetric graph G, the local variance will then be defined as:

$$v^*(y) = (1/n)\sum_{i=1}^{i=n}(y_i - m_i^*)^2, \quad \text{with:} \quad m_i^* = (1/n_i)\sum_{k=1}^{k=n_i} m_{ik} y_k$$

It is the average of the adjacent values of vertex i.

Note that if G is a complete graph (all pairs (i,i') are joined by an edge), *v*(y)* is nothing but *v(y)*, the classical empirical variance. When the observations are distributed randomly on the graph, both *v*(y)* and *v(y)* are estimates of the variance of y.

The contiguity ratio (analogue to the contiguity ratio of Geary, 1954), is written:

$$c^*(y) = v^*(y) \,/\, v(y), \text{ or : } \quad c^*(y) = \mathbf{y'(I - N^{-1}M)'\,(\,I - N^{-1}M)\,y} \;/\; \mathbf{y'\,(I - (1/n)U)\,y}$$

2.3 *Local Principal Component Analysis*

The contiguity ratio can be generalized :

i) to different distances between vertices in the graph,

ii) to multivariate observations (both generalizations are dealt with in: Lebart, 1969).

This section is devoted to the second generalization: the analysis of sets of multivariate observations having an *a priori* graph structure. Such situation occurs frequently in geography, ecology, geology. The multivariate analogue of the local variance is now the local covariance matrix, whose elements *cov(j,j')* are given by (using the previously defined notation):

$$cov^*(y_j, y_{j'}) = (1/n)\sum_{i=1}^{i=n}(y_i - m_i^*)(y_{j'} - m_{j'}^*)^2$$

If **X** designates the (*n,p*) data matrix giving the values of the *p* variables for each of the *n* vertices of the graph described by its associated matrix **M**, the local covariance matrix can be written :

$$\mathbf{V^*} = (1/n)\,\mathbf{X'(\,I - N^{-1}M)'\,(\,I - N^{-1}M)\,X}$$

The diagonalization of the corresponding local correlation matrix (Local Principal Component Analysis) produces a description of the local correlations, which can be compared to the results of a classical PCA performed with the global correlation matrix (see: Aluja and Lebart, 1984). Comparisons between covariance or correlation matrices (local and global) is usually done through Procustean Analysis (see: Gower and Dijksterhuis, 2004).

If the graph is made of *k* disjoined complete subgraphs, **V*** coincide with the classical "within covariance matrix" used in linear discriminant analysis.

If the graph is complete (associated matrix = **U**, with $u_{ij} = 1$ for all i and j), then **V*** is the classical covariance matrix, and the matrix : **(I -** *(1/n)* **U)** is idempotent.

$$\mathbf{V^* = V} = (1/n)\,\mathbf{X'(\,I - (1/n)\,U)'\,(\,I - (1/n)\,U)\,X} = (1/n)\,\mathbf{X'\,(\,I - (1/n)\,U)\,X}$$

2.4. *Contiguity Analysis*

Let **u** be a vector defining a linear combination *u(i)* of the *p* variables for vertex i:

$$u(i) = \Sigma_j\, u_j\, y_{ij} = \mathbf{u'y}_i$$

The local variance of the artificial variable *u(i)* is then :

$$v^*(\mathbf{u}) = \mathbf{u'\,V^*\,u}$$

The contiguity coefficient of this linear combination can be written :

$$c^*(\mathbf{u}) = \mathbf{u'\,V^*\,u} \;/\; \mathbf{u'\,V\,u}$$

where **V** is the classical covariance matrix of vector **y**.

The search for **u** that minimizes c*(**u**) produces functions having the properties of "minimal contiguity": these functions are, in a sense, the linear combinations of variables the more continuously distributed on the graph.

Instead of assigning an observation to a specific class, (as it is done in classical discriminant analysis) these functions allows one to assign it in a specific area of the graph. Therefore, this technique (designated as Contiguity Analysis) can be use to discriminate between overlapping classes.

3. SOM maps and external associated graph

The self organizing maps (SOM maps) proposed by Kohonen (1981) aim at clustering a set of multivariate observations. The obtained clusters are displayed as the vertices of a rectangular (chessboard like) or octagonal graph. The distances between vertices on the graph are supposed to reflect, as much as possible, the distances between clusters in the initial space.

3.1 Principles of the algorithm :

The size of the graph, and consequently, the number of clusters are chosen *a priori* (for example: a square grid with 5 rows and 5 columns, leading to 25 clusters). The algorithm is very similar to the McQueen algorithm (1967) in its on line version, and to the k-means algorithm (Forgy, 1965) in its batch version. This technique can be sketched as follows. Let us consider n points in a p-dimensional space (rows of the (n, p) matrix **X**). At the outset, to each cluster k is assigned a provisional centre $\mathbf{C}_k$ with p components (e.g.: chosen at random, or among the first elements). For each step t, the element i(t) is assigned to its nearest provisional centre $\mathbf{C}_k(t)$. Such centre, together with its neighbours on the grid, is then modified according to the formula:

$$\mathbf{C}_k(t+1)=\mathbf{C}_k(t)+\varepsilon\,(t)\,(i(t)-\mathbf{C}_k(t))$$

In this formula, $\varepsilon(t)$ is an adaptation parameter ($0< \varepsilon < 1$) which is a (slowly) decreasing function of t, as those usually encountered in stochastic approximation algorithms. This process is reiterated, and eventually stabilizes, but the partition obtained generally depends on the initial choice of the centres. In the batch version of the algorithm, the centres are modified only after a complete pass of the data set.

3.2 Graph associated with a SOM maps

Figure 1 represent a stylised symmetric matrix (70, 70) $\mathbf{M}_0$ associated to a partition of n=70 elements in k=8 classes (or clusters). Rows and columns represent the same set of n elements (elements belonging to a same class of the partition form a subset of consecutive rows and columns). The graph consists in a series of 8 cliques. All the cells of the black diagonal sub-matrices contains the value 1. All the cells outside these diagonal sub-matrices contains the value 0 .

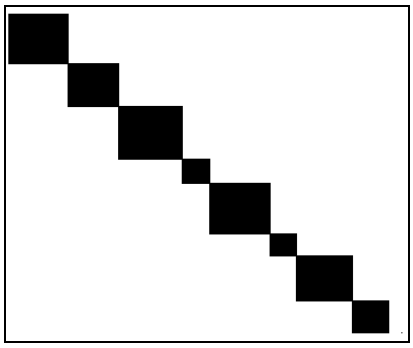

Figure 1 . Stylised incidence matrix $\mathbf{M}_0$ of the graph associated with a simple partition
(all the cells in the white [resp. black] areas contain the value 0 [resp. 1])

The 8 classes of the previous partition have been obtained through a SOM algorithm from a square 3 x 3 grid (with an empty class). The matrix of figure 1 does not take into account the topology of the grid: links between elements do exist only within clusters.

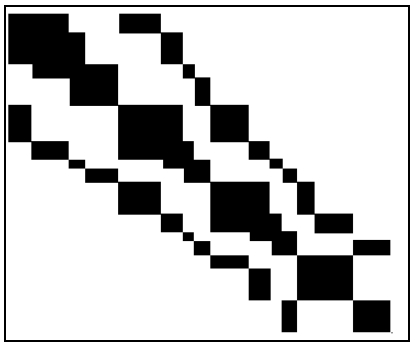

Figure 2 . Stylised incidence $\mathbf{M}_1$ matrix of the graph associated with a SOM map
(all the cells in the white [resp. black] areas contain the value 0 [resp. 1])

In figure 2, two elements i and j are linked ($m_{ij} = 1$) in the graph if they belong to a same cluster, or if they belong to contiguous clusters. Owing to the small size of the SOM grid (figure 3), the diagonal adjacency is not taken into account. (e.g.: elements belonging to cluster 7 are considered as contiguous to those of clusters 4 and 8, but not to the elements of cluster 5).

7	8	9
4	5	6
1	2	3

Figure 3. The *a priori* SOM grid.

Similarly to matrices M_0 and M_1, a matrix M_2 can be defined, that extends the definition of the edges of the graph to diagonal links. In the simple example of figure 3, the elements of cluster 7, for example, are considered as contiguous to the elements of clusters 4, 8, and 5.

4. Linear projectors onto the best SOM plane

The matrices M_0 and M_1, and M_2 can be easily obtained as a by product of the SOM algorithm.

4.1 Contiguity analysis using the graph G_0 the associated matrix of which is M_0:

In this case, the local variance coincide with the "within variance", and the result is a classical linear discriminant analysis of Fisher (LDA). In the plane spanned by the two first principal axes, the clusters are optimally located in the sense of the LDA criterion.

4.2 Contiguity analysis using the graphs G1 or G2 (associated matrices M_1, or M_2):

In those cases, the principal planes strive to reconstitute the positions of the clusters in the SOM map. In the initial p-dimensional space, the SOM map can be represented by the graph whose vertices are the centroids of the clusters. Those vertices are joined by an edge if the corresponding clusters are contiguous in the grid used in the algorithm. This graph in a high dimensional space will be partially or totally unfolded by the contiguity analysis. The following example will show the different phases of the procedure.

5. An example of application

An open-ended question has been included in a multinational survey conducted in seven countries (Japan, France, Germany, Italy, Nederland, United Kingdom, USA) in the late nineteen eighties (Hayashi *et al*., 1992).

The respondents were asked : *"What is the single most important thing in life for you?"* . This open question was followed by the probe: *"What other things are very important to you?"*.The illustrative example is limited to the British sample. The counts for the first phase of numeric coding are as follows: Out of *1043* responses, there are *13 669* occurrences (*tokens*), with *1 413* distinct words (*types*). When the words appearing at least *25* times are selected, there remain *9815* occurrences of these words, with *88* distinct words.

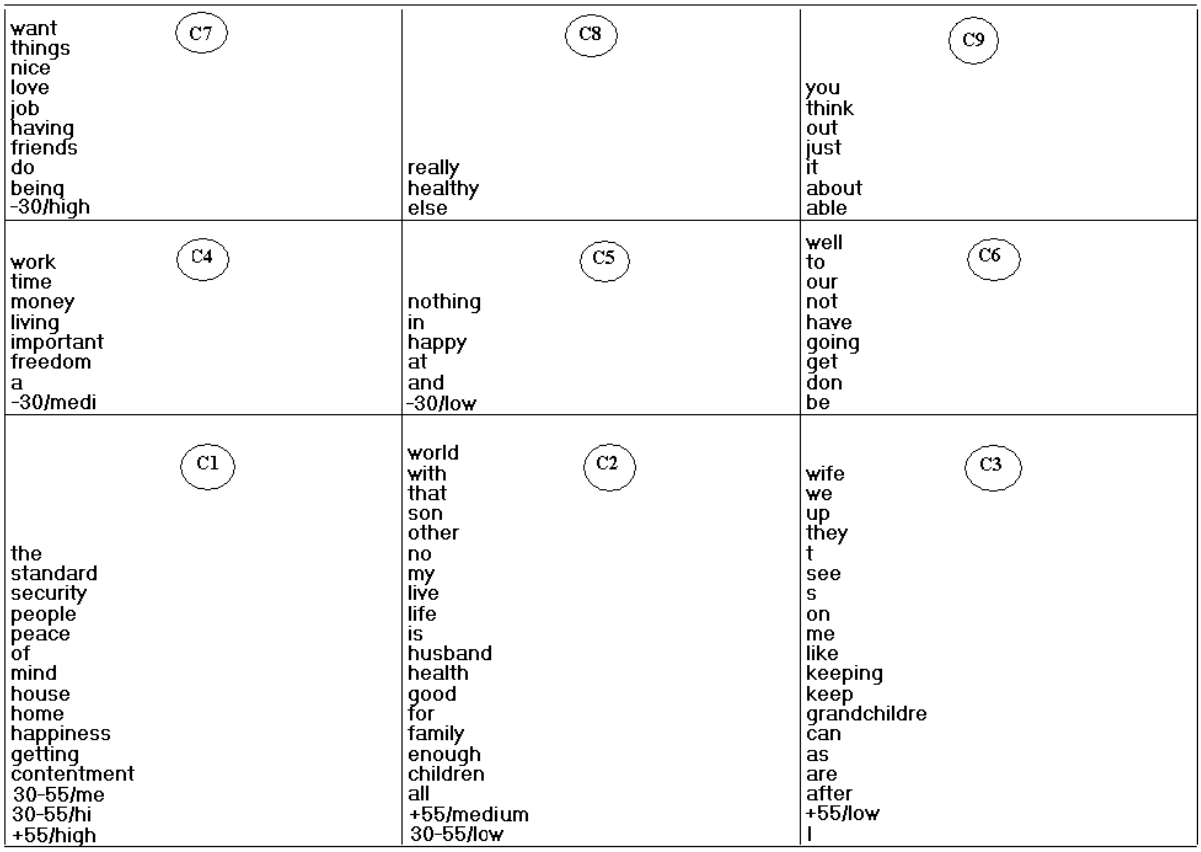

Figure 4. A (3 x 3) Kohonen map applied to the words used in the 1043 responses

In this example we focus on a partitioning of the sample into *9* categories, obtained by cross-tabulating age (*3* categories) with educational level (*3* categories). The 9 identifiers combine age categories (-30, 30-55, +55) with educational levels (low, medium, high).

Note that the SOM map (figure 4) provides a simultaneous representation of words and of categories of respondents. This is due to the fact that the input data are the coordinates provided by a correspondence analysis of the lexical contingency table cross-tabulating the words and the categories.

Figure 5 represents the plane spanned by the two first axes of the contiguity analysis using the matrix M_1. We can check that the graph describing the SOM map (the vertices of which C_1, C_2, …C_9 are the centroids of the elements of the corresponding cells of figure 4), is, in this particular case, a satisfactory representation of the initial map. The pattern of the nine centroids is similar to the original grid exemplified by figure 3.

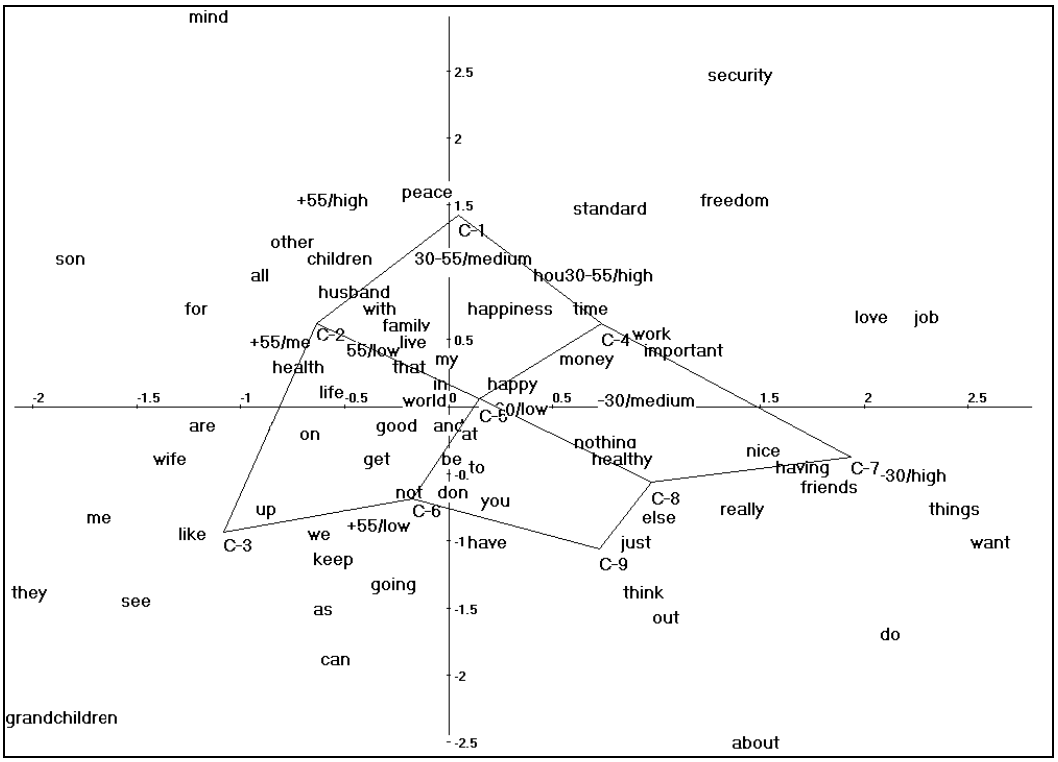

Figure 5. Principal plane of the contiguity analysis using matrix M1. The points C1, C2, …C9 represent the centroids of the 9 clusters derived from the SOM map.

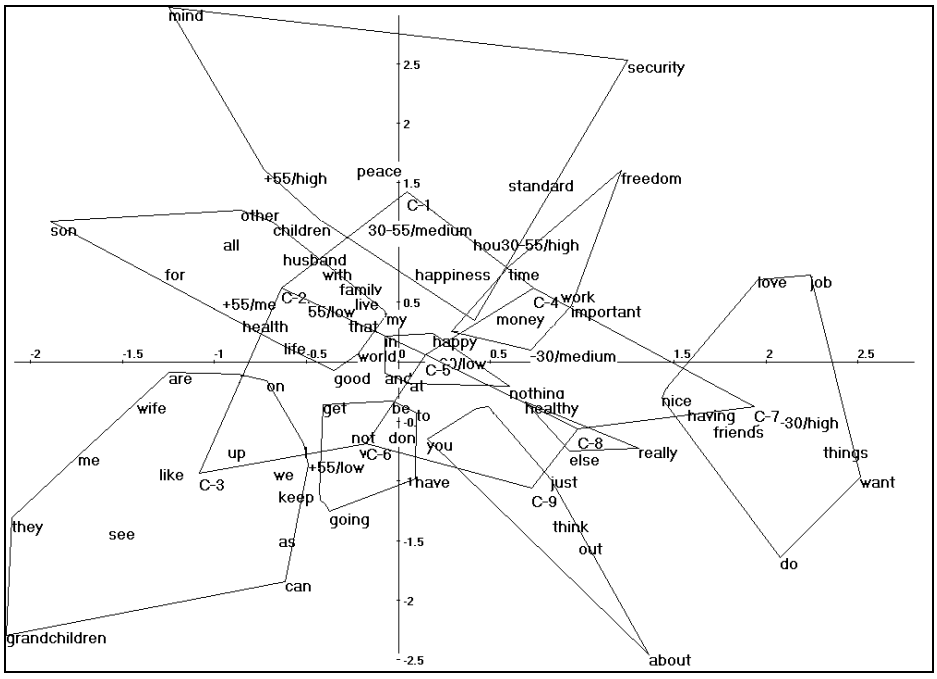

Figure 6. Principal plane of the contiguity analysis using matrix M1, with both the centroids of the 9 clusters and their convex hulls.

The background of figure 6 is identical to that of figure 5. It contains in addition the convex hulls of the nine clusters C1, C2, …, C9.. Each of those convex hulls correspond exactly (if we except some double or hidden points) to a cell of Figure 4. We note that these convex hulls are relatively well separated. In fact, figure 6 contains much more information than

figure 4, since we have now an idea of the shapes and sizes of the clusters, of the degree to which they overlap. We are now aware of their relative distances, and, another piece of information missing in figure 4, we can observe the configurations of elements within each cluster.

6. The assessment of Kohonen maps through partial bootstrap

We are provided at this stage with a tool allowing us to explore a continuous space. We can take advantage of having a projection onto a plane (and possibly onto a higher dimensional space, although the outputs are much more complicated in that case) to project the bootstrap replicates of the original data set. This can be done in the framework of a partial bootstrap procedure. In the context of principal axes techniques (such as singular values decomposition, principal component analysis, correspondence analysis, and also contiguity analysis), *Bootstrap* resampling techniques (see: Efron and Tibshirani, 1993) are used to produce confidence areas on two-dimensional displays. The bootstrap replication scheme allows one to draw confidence ellipses for both active elements (i.e.: elements participating in building principal axes) and supplementary elements (projected *a posteriori*).

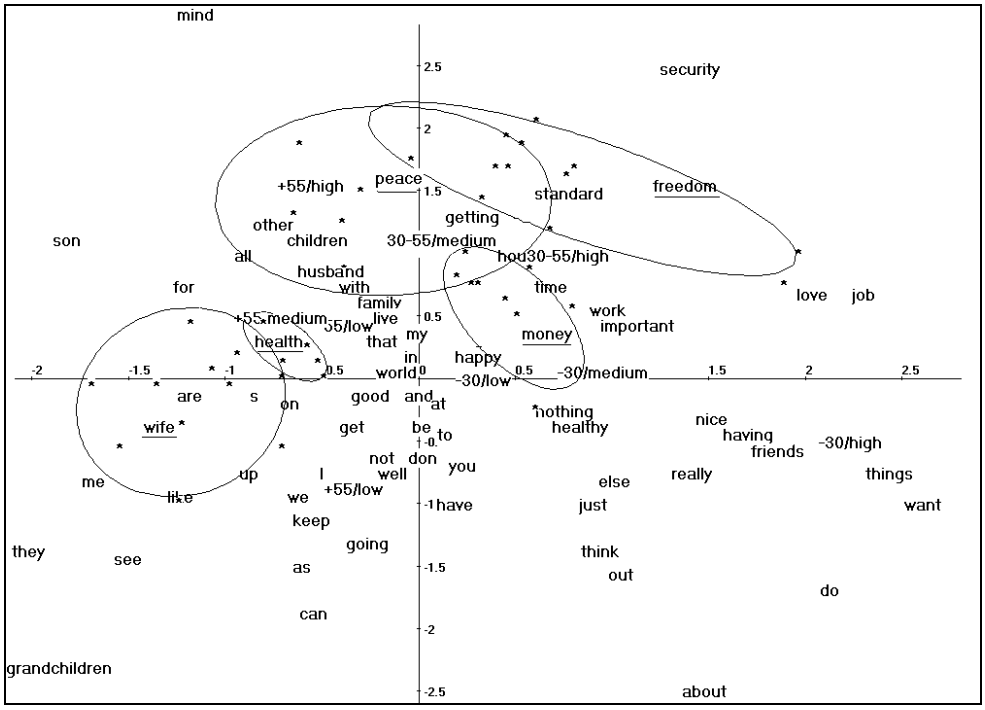

Figure 7. Bootstrap ellipses of confidence of the 5 words: *freedom, health, money, peace, wife* in the same principal contiguity plane as in figure 5 and 6.

In the example of the previous section, the words are the rows of a contingency table. The perturbation of such table under a bootstrap re-sampling procedure leads to new coordinates for the *replicated* rows. Without re-computing the whole contiguity analysis for each replicated sample (conservative procedure of total bootstrap), one can project the replicated rows as supplementary elements on a common reference space, exemplified above by figures 5 and 6. Always on that same space, figure 7 shows a sample of the replicates of five points (small stars visible around the words *freedom, health, money, peace, wife*) and the confidence ellipses supposed to contain approximately 90 % of these replicated points. Such procedure of partial bootstrap (see, e.g., Lebart, 2004) gives satisfactory estimates of the relative uncertainty about the location of points. Although the background of figures 6 and 7 are the same, it is preferable, to keep the results legible, to draw the confidence ellipses on a distinct

scattering diagram. It can be seen for instance that the words *freedom* and *money*, both belonging to cluster c4, have different behaviours with respect to the re-sampling variability. The location of *freedom* is much more fuzzy. That word could belong to other neighbouring clusters as well.

7. Conclusions

We have intended to immerse the self organizing maps, obtained through an algorithm often viewed as a black box, into an analytical framework (the linear algebra of contiguity analysis) and into an inferential setting as well (re-sampling techniques of bootstrap). That does not put into question the undeniable qualities of clarity and readability of the SOM maps. But it may perhaps help to assess the scientific status of these maps: like most exploratory tools, they may help to uncover rapidly and at low cost some features and patterns. However, they should undoubtedly be complemented by other statistical procedures if deeper interpretation is needed.

References

Aluja Banet T., Lebart L. (1984) Local and Partial Principal Component Analysis and Correspondence Analysis, *COMPSTAT Proceedings*, 113-118, Physica Verlag, Vienna.

Escofier B. (1989) Multiple correspondence analysis and neighboring relation. In : *Data Analysis, Learning Symbolic and Numeric Knowledge*. Diday E. (ed.), Nova Science Publishers, New York, p 55-62.

Efron B., Tibshirani R. J (1993)..: *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Forgy E. W. (1965) - Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *Biometric Society Meetings*, Riverside, California (Abstract in : *Biometrics* 21, 3, p 768).

Geary R.C. (1954) The Contiguity Ratio and Statistical Mapping, *The Incorporated Statistician*, 5, 115-145.

Gower J. C., Dijksterhuis G. B. (2004) Procustes Problem. Oxford Statistical Science Series, Oxford.

Hayashi C., Suzuki T., Sasaki M. (1992). *Data Analysis for Social Comparative Research: International Perspective*. North-Holland, Amsterdam.

Kohonen T. (1989). *Self-Organization and Associative Memory*. Springer Verlag, Berlin.

Lebart L. (1969) Analyse Statistique de la Contiguité, *Publications de l'ISUP*, XVIII, 81-112.

Lebart, L. (2000): Contiguity Analysis and Classification, In: W. Gaul, O. Opitz and M. Schader (Eds):*Data Analysis*. Springer,Berlin, 233--244.

Lebart L. (2004): Validation techniques in Text Mining. In: *Text Mining and its Application*, S. Sirmakensis (ed.), Springer Verlag, Berlin- Heidelberg, 169-178.

MacQueen J. B. (1967) - Some methods for classification and analysis of multivariate observations. *Proc. Symp. Math. Statist. and Probability (5th)*, Berkeley, **1**, p 281-297, Univ. of California. Press, Berkeley.