1

Multilevel Clustering for large Databases

Yves Lechevallier and Antonio Ciampi

INRIA-Rocquencourt,
78153 Le Chesnay CEDEX, France
Department of Epidemiology & Biostatistics,
McGill University, Montreal, P.Q., Canada

Abstract: Standard clustering methods do not handle truly large data sets and fail to take into account multi-level data structures. This work outlines an approach to clustering that integrates the Kohonen Self Organizing Map (SOM) with other clustering methods. Moreover, in order to take into account multi-level structures, a statistical model is proposed, in which a mixture of distributions may have mixing coefficients depending on higher-level variables. Thus, in a first step, the SOM provides a substantial data reduction, whereby a variety of ascending and divisive clustering algorithms become accessible. As a second step, statistical modelling provides both a direct means to treat multi-level structures and a framework for model-based clustering. The interplay of these two steps is illustrated on an example of nutritional data from a multi-center study on nutrition and cancer, known as EPIC.

Keywords and phrases: Clustering, Classification on very large databases, Data reduction

1.1 Introduction

Appropriate use of a clustering algorithm is often a useful first step in extracting knowledge from a data base. Clustering, in fact, leads to a *classification*, *i.e.* the identification of homogeneous and distinct subgroups in data [9] and [2], where the definition of *homogeneous* and *distinct* depends on the particular algorithm used : this is indeed a simple structure, which, in the absence of *a priori* knowledge about the multidimensional shape of the data, may be a reasonable starting point towards the discovery of richer, more complex structures.

In spite of the great wealth of clustering algorithms, the rapid accumulation of large data bases of increasing complexity poses a number of new problems that traditional algorithms are not equipped to address. One important feature of modern data collection is the ever increasing size of a typical data base: it

Table 1.1: French Center sample

Center	number	frequency
Ile-de-France	1201	24.75
Nord-Pas-de-Calais	452	9.32
Alsace-Lorraine	478	9.85
Rhone-Alpes	1018	20.98
Languedoc-Roussillon	625	12.88
Aquitaine	443	9.13
Bretagne-Pays-de-Loire	635	13.09

is not so unusual to work with data bases containing from a few thousands to a few millions of individuals and hundreds or thousands of variables. Now, most clustering algorithms of the traditional type are severely limited as to the number of individuals they can confortably handle (from a few undred to a few thousands). Another related feature is the multi-level nature of the data: typically a data base may be obtained from a multi-country, multi-centre study, so that individuals are nested into centres which are nested into countries. This is an example of an elementary, known structure in the data which should not be ignored when attempting to discover new, unknown structures.

This work arises from the participation of one of its authors to the EPIC project. EPIC is a multi-centre prospective cohort study designed to investigate the effect of dietary, metabolic and other life-style factors on the risk of cancer. The study started in 1990 and includes now 23 centres from 10 European countries. By now, dietary data are available on almost 500,000 subjects. Here we initiate a new methodological development towards the discovery of dietary patterns in the EPIC data base. We look for general dietary patterns, but taking into account, at the same time, geographical and socio-economic variation due to country and centres.

For simplicity, we consider only data from a subsample of the EPIC population consisting 4,852 of French women distributed in seven centres :

Also, we limit ourselves to an analysis of data from a 24-hour recall questionnaire concerning intake of sixteen food-groups. Thus, we will only discuss clustering for 2-level systems: subjects (first level) and centre (second level), in our case.

The approach we propose is based on two key ideas :

1) A preliminary data reduction using a Kohonen Self Organizing Map (SOM) is performed. As a result, the individual measurements are replaced by the means of the individual measurements over a relatively small number of *micro-regimens* corresponding to Kohonen neurons. The micro-regimens

can now be treated as new *cases* and the means of the original variables over micro-regimens as new *variables*. This *reduced* data set is now small enough to be treated by classical clustering algorithms. A further advantage of the Kohonen reduction is that the vector of means over the micro-regimens can safely be treated as multivariate normal, owing to the central limit theorem. This is a key property, in particular because it permits the definition of an appropriate dissimilarity measure between micro-regimens.

2) The multilevel feature of the problem is treated by a statistical model wich assumes a mixture of distributions, each distribution representing, in our example, a *regimen* or *dietary pattern*. Although more complex dependencies can be modeled, here we will assume that the centres only affect the mixing coefficients, and not the parameters of the distributions. Thus we look for general dietary patterns assuming that centers differ from each other only in the distribution of the local population across the general dietary patterns.

While the idea of a preliminary Kohonen reduction followed by the application of a classical clustering algorithm is not entirely new [12], [1] and [14], this work differs from previous attempts in several respects the most important of which are :

a) the Kohonen chart is trained by an the initialization based on principal component analysis;

b) the choice of clustering algorithm is guided by the multilevel aspect of the problem at hand;

c) the clustering algorithm is based on a statistical model.

Thus this work continues the author's research program which aims to develop data analytic strategies integrating KDDM and classical data analysis methods [6] and [7].

1.2 Data Reduction by Kohonen SOM's

We consider p measurements performed on n subjects grouped in C classes, $\{G_c, c = 1, \ldots, C\}$. We denote these measurements by $(G^{(i)}, x^{(i)}), i = 1, \ldots, n$, where for the i^{-th} subject $G^{(i)}$ denotes the class (the centre, in our example), and $x^{(i)}$ the p-vector of measurements (the 16 food-group intake variables); or, in matrix form, $\mathbf{D} = [\mathbf{G}|\mathbf{X}]$.

In this section we describe the first step of the proposed approach, which consists in reducing the $n \times p$ matrix $\mathbf{X}$ to a $m \times p$ matrix, $m \ll n$. To do this, we first pass the data matrix $\mathbf{X}$ through a Kohonen SOM consisting of m units (neurons) disposed in a rectangular sheet with connections along two perpendicular axis.

1.2.1 Kohonen SOM's and PCA initialization

We recall that in a Kohonen SOM the neurons of the rectangular sheet are associated to a grid of prototypes in the p-dimensional space which represents the row-vectors of the data matrix: the sheet is supposed to represent the grid with a minimum distortion, so that a SOM can be seen as a non-linear version of classical data reduction techniques such as a Principal Component Analysis (PCA). In order to specify a SOM, one needs to specify initial values of the sheet's connection weights and of the position of the prototypes. Then, the data points are repeatedly sent through the SOM, each passage causing an update of both the connection weights and the position of the prototypes, *i.e* an alteration of both the sheet in 2-dimensional space and the grid in p-dimensional space. Normally, this process converges, in that the changes at each passage become negligible.

In the original approach, initial weights were chosen at random; however, as the efficacy of the algorithms crucially depends on the initialization, much effort has been devoted to improving this first step. The distinguishing feature of our construction consists in designing the sheet with the help of the results of PCA performed on $\mathbf{X}$. It is advantageous to choose the dimensions of the grid, a and b, $(m = ab)$, such that :

$$\frac{a}{b} = \frac{\sqrt{\lambda_1}}{\sqrt{\lambda_2}}$$

where λ_1 and λ_2 denote the first and second eigenvalues of the PCA, see figure 1.1. Also, the initial connection weights and position of the prototypes are obtained from the two first eigenvectors of the PCA. The details are described in [8], where it is also shown that PCA initialization presents substantial practical advantages over several alternative approaches.

1.2.2 Binning of the original data matrix using a Kohonen Map

As a result of the training process, the SOM associates to each subject a unique neuron-prototype pair, which we shall refer to as *micro-regimen.* Each micro-regimen , B_r, $r = 1, \ldots, m$, can be considered as a *bin* in which similar individuals are grouped. We shall denote by n_r the number of subjects in B_r and by $n_{r,c}$ the number of subjects in $B_r \cap G_c$. Let also $\bar{x}_r$ and $\bar{x}_r^{(c)}$ denote the vectors of the means of $x^{(i)}$ taken over B_r and over $B_r \cap G_c$ respectively. Figure 1.2 gives a graphical representation of the bins [10] : in each bind the dot is proportional to bin size and the graph is a profile of the input variables.

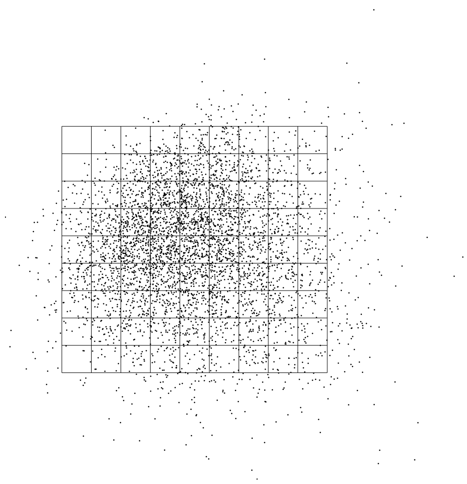

Figure 1.1: Initialization by PCA

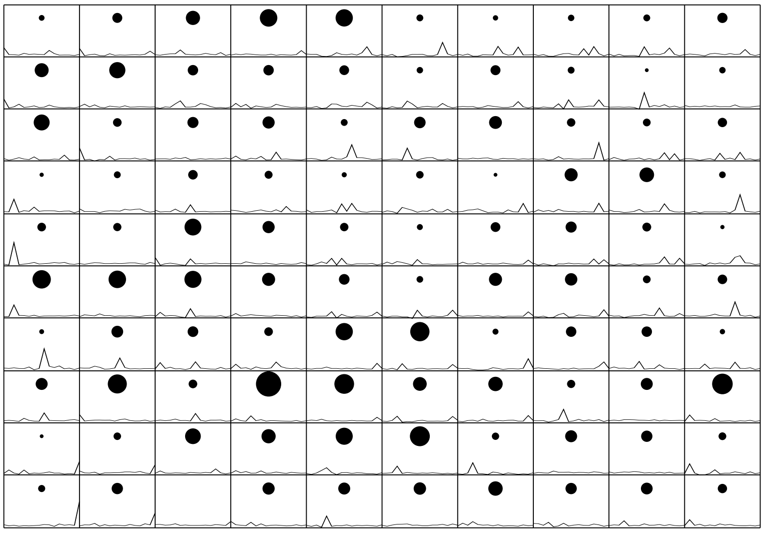

Figure 1.2: Kohonen map

Already at this stage, an exploratory analysis of the two-way table $\{n_{r,c}; r = 1, \ldots, m,\ c = 1, \ldots, C\}$, would be instructive: *e.g.* Correspondence Analysis (CA) of the table, ordering its rows and columns according to the factor scores and eventually clustering rows and columns, is likely to shed some light on the relationship between centers and micro-regimens.

Our goal, however, is to look for macro-regimens, (dietary patterns in our example), by clustering micro-regimens. To proceed further, we assume here that the expected value of x and its variance-covariance matrix may depend on the micro-regimens but not on the centers. It follows that if $n_{r,c}$ is large enough, then, by the central limit theorem, $\bar{x}_r^{(c)}$ is approximately multivariate normal $\mathcal{N}_p(\mu_r, \frac{1}{n_{r,c}}\Sigma_r)$ and the maximum likelihood estimate of μ_r and Σ_r are :

$$\bar{x}_r = \frac{1}{n_r}\sum_{i \in B_r} x^{(i)} \quad \text{and} \quad V_r = \frac{1}{n_r}\sum_{i \in B_r} (x^{(i)} - \bar{x}_r)^T (x^{(i)} - \bar{x}_r)$$

1.2.3 Dissimilarity for micro-regimens

From these consideration, a natural definition for a dissimilarity between two *bins* B_r and B_s follows. This is the likelihood ratio statistic (LRS) comparing the hypothesis that $\bar{x}_r^{(c)}$ and $\bar{x}_s^{(c)}$ have different distributions with the hypothesis that they have the same distribution.

$$d(B_r, B_s) = 2\sum_{s=1}^{C} log[\frac{\mathcal{N}_p(\bar{x}_r^{(c)}|\bar{x}_r, \frac{1}{n_{r,c}}V_r)\mathcal{N}_p(\bar{x}_s^{(c)}|\bar{x}_s, \frac{1}{n_{s,c}}V_s)}{\mathcal{N}_p(\bar{x}_r^{(c)}|\bar{x}_{r\cup s}, \frac{1}{n_{r,c}}V_{r\cup s})\mathcal{N}_p(\bar{x}_s^{(c)}|\bar{x}_{r\cup s}, \frac{1}{n_{s,c}}V_{r\cup s})}] \tag{1.1}$$

where $\mathcal{N}_p(.|\mu, \Sigma)$ is the density function of a multivariate normal $\mathcal{N}_p(\mu, \Sigma_r)$ and

$$\bar{x}_{r\cup s} = \frac{n_r\bar{x}_r + n_s\bar{x}_s}{n_r + n_s} \text{ and} \tag{1.2}$$

$$V_{r\cup s} = \frac{1}{n_r + n_s}[n_rV_r + n_sV_s + n_r.n_s(\bar{x}_r - \bar{x}_s)(\bar{x}_r - \bar{x}_s)^t]. \tag{1.3}$$

1.3 Clustering multi-level systems

While the dissimilarity of equations (1) and (2) is very natural in our context, other ones can be usefully defined, for example those proposed in the symbolic data analysis literature [4]. Once the choice of dissimilarity has been made, several standard algorithms can be applied to the bins. Since the number of bins ($m \ll n$) may be chosen to be relatively small, a panoply of ascending approaches becomes accessible. Moreover, several dissimilarity-based divisive approaches are available. Among these, some conceptual clustering algorithms [5] seem particularly promising as the one used in this work, see next section.

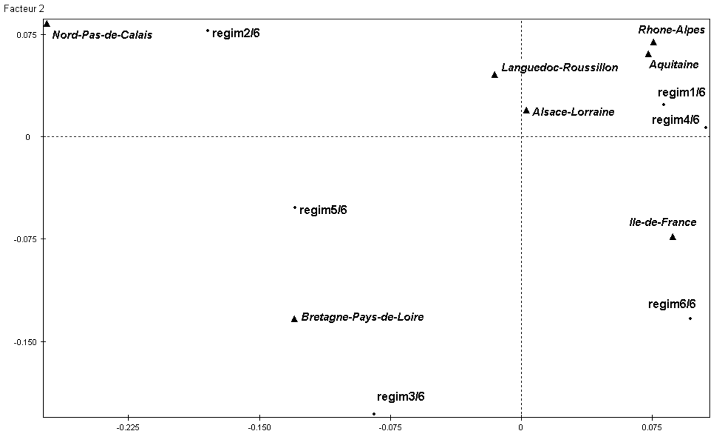

Figure 1.3: Relation between center and regimens

Suppose now that a clustering algorithm has been deployed. As a result, the m micro- regimens are grouped to produce $k \ll m$ macro-regimens. Furthermore, the 2-way table $\{m_{i,c}; i = 1, \ldots, k,\ c = 1, \ldots, C\}$ obtained by crossing centers with macro-regimens can be analyzed by CA as outlined in the previous section. Finally, proportions of subjects following different regimens in each centre would usefully summarize local characteristics, while a description of the clusters would give insight on the nature of the macro-regimens found in the general population.

1.3.1 A two level statistical model

A statistical model may now be proposed. This can be useful for efficiently extracting information from the data, as it suggests a family of model-based clustering algorithms which explicitly account for the multi-level structure of the data. For a two-level system we suppose that the reduced data vector $\bar{x}_r^{(c)}$ has as distribution a mixture of k multivariate normal distributions, each corresponding to a macro-regimen, or *dietary pattern* in our example. Thus the density can be written as :

$$f(\bar{x}_r^{(c)}) = \sum_{i=1}^{k} \alpha_i(G_c)\mathcal{N}_p(\bar{x}_r^{(c)}|\mu_i, \Delta_i) \tag{1.4}$$

A more complex model would include dependence of the μ's and the Δ's on c. The interest of such a model is limited, although it could eventually be used to check the adequacy of the one we propose. The simpler model is of greater interest, especially for our dietary data example, because it allows identification of general patterns which are to be found in all centres *albeit* in different proportions. For instance we may expect that the *mediterranean diet* is not an exclusive characteristics of mediterranean regions; though more frequently encountered in these regions, it can be chosen as a way of eating normally, perhaps for health reasons, by people living in all areas of France, and indeed, of Europe.

It is easy to see how this 2-level model can be generalized to three- and multi-level systems by introducing, for example, a country level and treating centers-within-country by random effects. This, however, will not be pursued here.

1.3.2 Estimating parameters by the EM algorithm

In many situations, a reasonable description of the data is amply sufficient. Then the statistical model of equation (3) serves as useful guidance, but the exploratory approach outlined above is all that is needed: indeed, it produces both a reasonable guess for the number of clusters and rough estimates of means, variance-covariance matrices and mixing coefficients. On the other hand, when more precise estimates are desired, these rough ones can be used to initialize an iterative algorithm for maximum likelihood estimation. Here, as we are dealing with a mixture model, the EM algorithm seems an appropriate choice, with the dependence of the mixing coefficients on centre introducing only a minor additional complication.

The EM is applied as follows :

a) The complete data are : $(\bar{x}_r^{(c)}, \rho(r))$, where $\rho(r)$ is the (actually unknown) regimen to which the r-th microregimen belongs;

b) The likelihood of the complete data is :

$$l = logL = \sum_{c=1}^{C}\sum_{r=1}^{m}\sum_{i=1}^{k} log[\alpha_i^{(t)}(G_c) + \mathcal{N}_p(\bar{x}_r^{(c)}|\mu_i, \tfrac{1}{n_{r,c}}\Delta_i)]$$

c) At step t, let :

$$p^{(t)}(c|i, \bar{x}_r^{(c)}) = \frac{\alpha_i^{(t)}(G_c)\mathcal{N}_p(\bar{x}_r^{(c)}|\mu_i^t, \tfrac{1}{n_{r,c}}\Delta_i^t)}{\sum_{j=1}^{k}\alpha_j^{(t)}(G_c)\mathcal{N}_p(\bar{x}_r^{(c)}|\mu_j^t, \tfrac{1}{n_{r,c}}\Delta_j^t)}$$

Then the iteration equations from the EM approach can be shown to be :

$$\begin{aligned}
\mu_i^{(t)} &= \frac{1}{n}\sum_{c=1}^{C}\sum_{r=1}^{m} p^{(t-1)}(c|i, \bar{x}_r^{(c)})).\bar{x}_r^{(c)} \\
\Delta_i^{(t)} &= \frac{1}{n}\sum_{c=1}^{C}\sum_{r=1}^{m} p^{(t-1)}(c|i, m^{(r,c)})(\bar{x}_r^{(c)}) - \mu_i^{(t-1)})^T(\bar{x}_r^{(c)}) - \mu_i^{(t-1)})
\end{aligned}$$

Table 1.2: Proportion of the 6 regimens: overall and by centre

Regimens	Overall	Alsace -Lorraine	Aquitaine	Bretagne Loire	Ile-de -France	Languedoc -Roussillon	Nord -Pas -de-Calais	Rhone -Alpes
regim 1	0,56	0,58	0,59	0,49	0,58	0,54	0,46	0,61
regim 2	0,19	0,18	0,18	0,20	0,14	0,21	0,28	0,18
regim 3	0,08	0,08	0,07	0,12	0,09	0,08	0,08	0,06
regim 4	0,03	0,02	0,04	0,03	0,03	0,04	0,02	0,03
regim 5	0,10	0,10	0,08	0,11	0,10	0,08	0,13	0,09
regim 6	0,04	0,04	0,04	0,05	0,05	0,04	0,03	0,04

$$p^{(t)}(c|i, \bar{x}_r^{(c)}) \quad = \quad \frac{1}{n_c} \sum_{s=1}^{m} p^{(t-1)}(c|i, \bar{x}_s^{(c)}))$$

1.4 Extracting dietary patterns from the nutritional data

We return now to the subset of the EPIC data base describing dietary habits of 4,852 French women. Figure 1.2 summarises the Kohonen SOM analysis of the data based on a 10×10 sheet. Since one bin is empty, 99 distinct regimens were identified. Both a standard ascending algorithm [12] and a conceptual clustering algorithm [5] applied to the micro-regimens, suggest 4, 6 or 9 classes or dietary patterns. The results of the 6-class analysis are summarised in Figure 1.4, which shows the first factorial plane of the CA representing the relationship between centres and dietary pattern; Figure 1.4, which shows the Zoom Star graphs [13] of the eight most discriminating variables describing dietary patterns; and Table 1.4 which gives a rough estimate of the proportions of subjects following the six dietary patterns, overall and by centre.

An example of interpretation is as follows: regimen 1 is characterized by high consumption of meat and vegetables; regimen 2 by high soups and low vegetable consumption; regimen 3 by high fish and low meat consumption (respectively 13% and 3% of the total weight of food intake); regimen 4 by high meat and low fish consumption; regimen 5 by high alcohol and meat consumption; and regimen 6 by high consumption of dairy products, eggs and vegetables and low consumption of fish, alcoholic beverage and legumes. Also, the Nord-Pas-de-Calais region is positively associated to regimen 2 and 5 and negatively to regimen 1; similarly, there is a positive association of Bretagne-Pays-de-Loire with regimen 3 and a negative association with regimen 1; and finally, Rhone-Alpes is positively associated to regimen 1.

Aknowledgments :

We wish to thank Dr. F. Clavel for having shared the French data with us, and Dr. E. Riboli for general assistance in becoming familiar with EPIC.

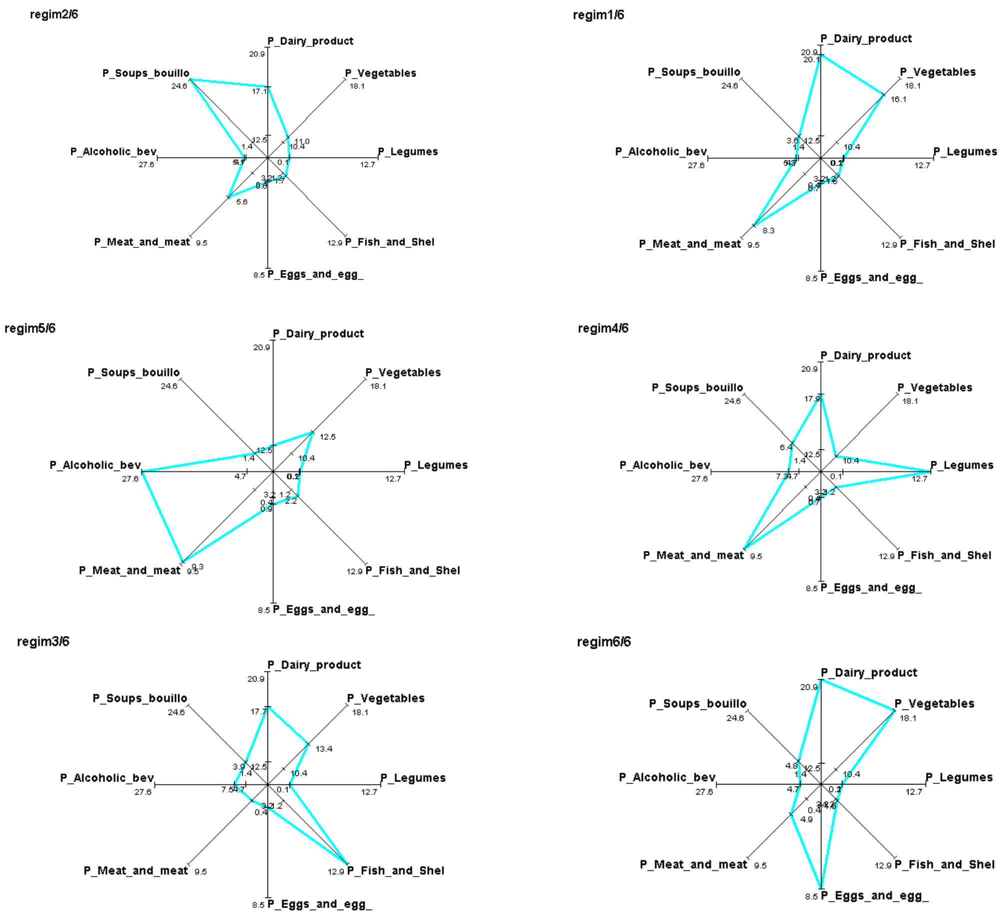

Figure 1.4: The 6 regimens by Zoom Stars

Bibliography

[1] Ambroise, C., Seže, G., Badran, F., Thiria, S.: Hierarchical clustering of Self-Organizing Maps for cloud classification. *Neurocomputing*, *30*, (2000) 47–52.

[2] Bock, H. H.: Classification and clustering : Problems for the future. In: E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, B. Burtschy (eds.): *New Approaches in Classification and Data Analysis.* Springer, Heidelberg (1993), 3–24.

[3] Bock, H. H.: Clustering and neural networks. In: A. Rizzi, M. Vichi, and H.-H. Bock (Eds.): *Advances in Data Science and Classification.* Springer, Heidelberg (1998), 265–278.

[4] Bock, H. H., Diday, E. (Eds.): *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data.* Studies in Classification, Data Analysis and Knowledge Organization, Springer, Heidelberg (1999).

[5] Chavent, M.: A monothetic clustering algorithm. *Pattern Recognition Letters*, *19*, (1998) 989–996.

[6] Ciampi, A., Lechevallier, Y.: Designing neural networks from statistical models : A new approach to data exploration. Proceedings of the 1^{st} International Conference on Knowledge Discovery and Data Mining. AAAI press, Menlo Park, Capp. (1995) 45–50.

[7] Ciampi, A., Lechevallier, Y.: Statistical Models as Building Blocks of Neural Networks. *Communications in Statistics*, *26(4)*, (1997) 991-1009.

[8] Elemento, O.: Apport de l'analyse en composantes principales pour l'initialisation et la validation de cartes de Kohonen. *Septièmes Journées de la Société Francophone de Classification*, Nancy (1999).

[9] Gordon, A. D.: *Classification : Methods for the Exploratory Analysis of Multivariate Data.* Chapman & Hall, London (1981).

[10] Hébrail, G., Debregeas, A.: Interactive interpretation of Kohonen maps applied to curves. Proceedings of the 4^{th} International Conference on Knowledge Discovery and Data Mining. AAAI press, Menlo Park (1998) 179–183.

[11] Kohonen, T.: *Self-Organizing Maps*. Springer, New York (1997).

[12] Murthag, F.: Interpreting the Kohonen self-organizing feature map using contiguity-constrained clustering. *Patterns Recognition Letters*, *16*, (1995) 399–408.

[13] Noirhomme-Fraiture, M., Rouard, M.: Representation of Sub-Populations and Correlation with Zoom Star. *Proceedings of NNTS'98*, Sorrento (1998).

[14] Thiria, S., Lechevallier, Y., Gascuel, O., Canu, S.: *Statistique et méthodes neuronales*. Dunod, Paris, (1997).