UNIVERSITE PARIS 1 PANTHEON SORBONNE

CENTRE DE RECHERCHE

S.A.M.O.S

STATISTIQUE APPLIQUEE ET MODELISATION STOCHASTIQUE

The Kohonen algorithm: a powerful tool for analyzing and representing multidimensional quantitative and qualitative data

M. Cottrell, P. Rousset

Prépublication du SAMOS n° 76
Mai 1997

90, rue de Tolbiac - 75634 PARIS CEDEX 13

THE KOHONEN ALGORITHM : A POWERFUL TOOL FOR ANALYZING AND REPRESENTING MULTIDIMENSIONAL QUANTITATIVE AND QUALITATIVE DATA

Marie Cottrell, Patrick Rousset

SAMOS, Université Paris 1
90, rue de Tolbiac
75634 PARIS Cedex 13
FRANCE
Tél et Fax : (33 1) 40 77 19 22
E-Mail : cottrell, rousset@univ-paris1.fr

Topics : Using a Kohonen algorithm to analyze and visualize multidimensional data involving quantitative and qualitative variables.
Topic No 7 : Methodology for Data Analysis, Task Selection and Nets Design.

Keywords : Kohonen maps, classification, multidimensional data analysis, general non linear models, neural networks.

Preferred presentation : Oral

Corresponding author : Marie Cottrell
SAMOS, Université Paris 1, 90, rue de Tolbiac,
75634 PARIS Cedex 13, FRANCE
Tél et Fax : (33 1) 40 77 19 22
E-Mail : cottrell@univ-paris1.fr

THE KOHONEN ALGORITHM :
A POWERFUL TOOL FOR ANALYSING AND REPRESENTING MULTIDIMENSIONAL QUANTITATIVE AND QUALITATIVE DATA

Marie Cottrell, Patrick Rousset

SAMOS, Université Paris 1
90, rue de Tolbiac
75634 PARIS Cedex 13
FRANCE
Tél et Fax : (33 1) 40 77 19 22
E-Mail : cottrell, rousset@univ-paris1.fr

Keywords : Kohonen maps, classification, multidimensional data analysis, general non linear models, neural networks.

Abstract

The simultaneous analysis of quantitative and qualitative variables is not an easy task in general. When a linear model is appropriate, the Generalized Linear Models are commonly used with success. But when the intrinsic structure of the data is not at all linear, they give very poor and confusing results. In this paper, we extensively study how to use the (non linear) Kohonen maps to solve some of the interesting problems which are encountered in data analysis : how to realize a rapid and robust classification based on the quantitative variables, how to visualize the classes, their differences and homogeneity, how to cross the classification with the remaining qualitative variables to interpret the classification and put in evidence the most important explanatory variables.

1 Introduction

The aim of this paper is the analysis of multidimensional data, involving quantitative (continuous) variables and qualitative (nominal, ordinal) variables. When the standart linear statistical methods are not appropriate, due to the intrinsic structure of the observations, one can try to use neural models because they are highly non linear. One can use for example a Multilayer Perceptron which admit at the same time quantitative and qualitative variables as inputs. But in that case, it is not very easy to interpret the model, to find the most relevant variables, to propose a relevant typololy of the observations from the results. Even if many advances have been realized from the « black box » epoch, to better choose the architecture of the network, to prune the non significant connections, to propose the extraction of rules, and so on, this kind of models is more appropriate for regression analysis or short term forecasting than for interpretation and visualization of the variables which describe the data. We can refer to the three last proceedings of the IWANN Conference to find many examples of these affirmations.

On the other hand, the Kohonen algorithm ([14], [15], [3], [5]) is widely used for data analysis ([2], [4], [6]), but its performances are not exhaustively used until now. After previous work using the Kohonen algorithm to realize long term forecasting by combining prevision and classification, [7], we extend the field of applications of this algorithm to a more general setting.

Let us give some notations : we consider a set of N observations, where each individual is described by p quantitative real valued variables and q qualitative variables. The main tool is a Kohonen network, generally a two-dimensional grid, with n by n units, but the method can be used with any topological organization of the Kohonen network. After learning, each unit i is represented in the R^p space by its weight vector C_i (or *code* vector). We do not address here the delicate problem of the learning of the code vectors which is supposed to be successfully realized from the N observations restricted to their p quantitative variables.

Then each observation is classified by a nearest neighbor method, (in R^p): observation k belongs to class i if and only if the code vector C_i is the closest among all the code vectors. The distance in R^p is the Euclidean distance in general, but it can be chosen in another way according to the application.

With respect to any other classification method, the main characteristic of the Kohonen classification is the conservation of the topology: after learning, « close » observations are associated to the same class or to « close » classes according to the definition of the neighborhood in the Kohonen network. This feature allows to consider the resulting classification as a good starting point to further developments as stated below.

In the next sections, we try to give some answers to the following problems : how to realize a rapid and robust classification based on the quantitative variables (Section 2), how to visualize the classes, their differences and homogeneities (Section 3), how to cross the classification with the remaining qualitative variables to interpret the classification and put in evidence the most important explanatory variables (Section 4), how to better understand the relations between a class and its neighbors (Section 5).

In order to explain the different aspects, we present the same example throughout the paper. The data come from the curves of daily half-hour electrical consumption [7] and have been transformed (for confidentiality purpose). We will consider « daily curves », with 12 quantitative dimensions (one « observation » each two hours) and 2 qualitative variables (the day and the month). The topology of the Kohonen network is cylindrical, in order to take into account the annual periodicity. All the graphes in the paper are related to this example. All the subroutines are developped with the SAS software, version 6.11. Actually these techniques have been applied to many other kinds of data (French unemployment, [11], Canadian individual consumption, [12], Ile-de France economic data, [13], and so on.

2 Two-levels classification

As mentioned above, the first and raw result we get after learning is a classification of the N observations into $P = n \times n$ classes. Eventually some classes can be empty. At the same time, we get the code vectors (the weights associated to each unit). See in Fig. 1, a standart representation of the code vectors inside their own unit.

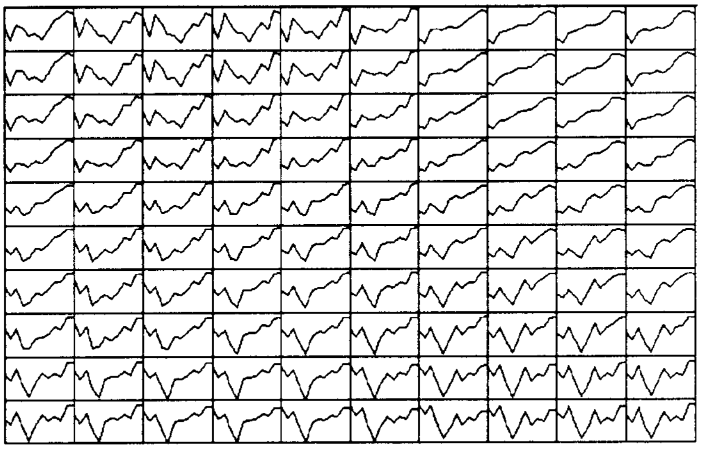

Fig. 1 : *There are 100 units in a grid. In each cell, the final 12-dimensional code vector (or weight vector) is drawn. Note that neighboring cells correspond to similar vectors.*

The choice of the number P of units is arbitrary, and there does not exist any method to better choose the size of the network. We can guess that the « relevant » number of classes could often be smaller than P (which is commonly equal to 100). It is also difficult to give relevant interpretation of a too large number of classes. So we propose to reduce the number of classes by means of a hierarchical classification [1] of the P code vectors using the Ward distance for example. As the code vectors are already organized across the grid, a standart method appears to be relevant.

In this way, we define two embedded classifications, and can distinguish the classes (Kohonen classes or « micro-classes ») and the « macro-classes » which group together some of the « micro-classes ». To make visible this two-levels classification, we affect to each « macro-class » some colour or trame or grey level (here). See in Fig.2, a representation of the « micro-classes » grouped together to constitute 10 « macro-classes ».

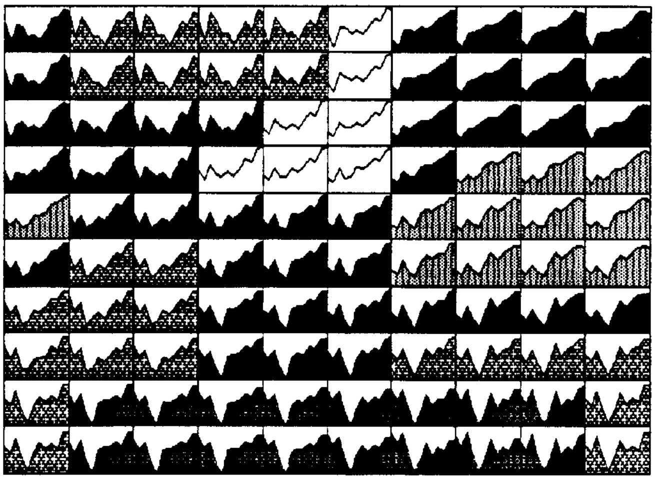

Fig. 2 : *The 10 « macro-classes » resulting of the hierarchical classification of the 100 code vectors are superposed on the grid.*

The advantage of this double classification is the possibility to analyze the data set at a « macro » level where general features emerge and at a « micro » level to determine the characteristics of more precise phenomena and especially the paths to go from one class to another one.

In the applications that we treated, the « macros classes » create connected areas in the grid (one « macro class » which contain some « micro class » contain also a neighbor of this class). This remark is very attractive because it confirms the topological properties of the Kohonen maps. Nevertheless, in some cases, it is possible to find a « macro class » split into two pieces. In that case, one can guess that the data set is folded over and try to control it by studying the distortion.

3. Analysis of the classes : discrimination and homogeneity

After the classes are defined, any standart statistical criterion can be computed to measure the inter- and intra- classes variances. See [1] for example for references. As a complement, we propose graphical methods to visualize them.

3.1 Discrimination

As suggested in the previous Section, it is important to have a better representation of the map geometry. The code vectors draw a broad outline of a N-dimensional surface with irregular distances between classes, but as in Fig. 2, the standart representation is a grid with a regular disposition of the units. This can produce confusions in the interpretation and it is valuable to visualize the distances between classes, [9]. This visualization avoids misleading interpretations and gives an idea of the discrimination between classes.

We use the method proposed by [9] : each unit is represented by an octagon. The bigger it is, the closer the unit is to its neighbors. So the clusters appear to be regions in which octagons tend to be big and frontiers are regions largely unshaded. See an example in Fig. 3.

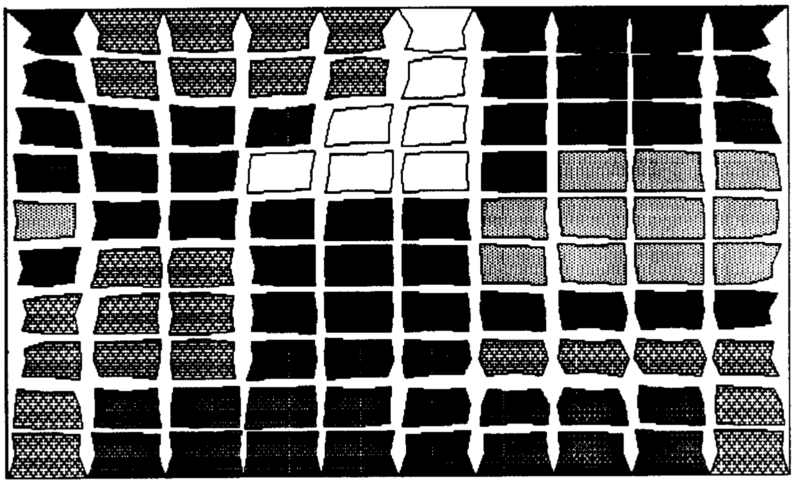

Fig. 3 : *Representation of the distances : actually the gaps coincide with the frontiers of the « macro-classes ».*

We can observe that in general the « macro-classes » boundaries coincide with the most important distances between classes, and that confirms the pertinence of the second level classification. On the contrary, if a boundary occurs between two classes with small distance, that means that the second level classification splits a large group into two groups and that the path from one to the other is continous. It indicates that perhaps we could consider a hierarchical classification with fewer classes.

3.2 Contents of the classes, homogeneity

Another question is how to put in evidence the intra-classes dispersion. This is related to the problem of the outliers or of the existence of a small typical group different from the rest of the data. We try to presente a visual tool to decide which observation could be deleted in the learning phase, (because it is too far from the other observations or can be erroneous,...) and also how it would have been classified after learning.

See in Fig. 4 the representation of all the observations inside their own class. We can immediately see in which units the dispersion is large with respect to the others, which observations could be deleted or examined separately and which unit gathers together in fact two different populations.

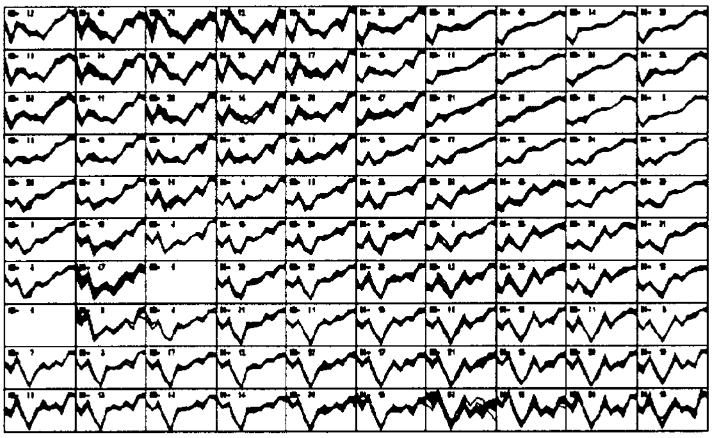

Fig. 4 : *The contents of the classes and the numbers of their elements*

4 Crossing the classification with a qualitative variable

In this part, the goal is to answer to these questions : which are the observations of this class, does there exist a characteristic common to the neighboring classes, can we qualify a group of classes by a qualitative variable?. A first method can be to extract the observations of a given class and analyse them with a statistical software, by computing means, variances for the quantitative variables (used for the classification) and frequencies for the qualitative variables. That gives an answer to the first question, but we lose the neighborhood properties of the Kohonen map.

In order to complete the description, we study the repartition of each qualitative variable inside each class. Let be Q a qualitative variable with K modalities. In each cell of the Kohonen map, we draw a frequency pie, where each modality is represented by a grey level occupying an area proportional to its frequency in the corresponding class. See in Fig. 5 such an example of a frequency pie.

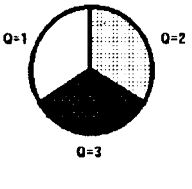

Fig. 5 : *A frequency pie, when there are 3 modalities with frequencies 1/3.*

So in this way, by representing the frequencies of each modality across the map, we make clear the continuity between some classes as well as the breakings. See in Fig.

6 an example of this visualization technique. Here the qualitative variable is the day, with 3 levels : Sundays (black), Saturday (grey) and week-days (white).

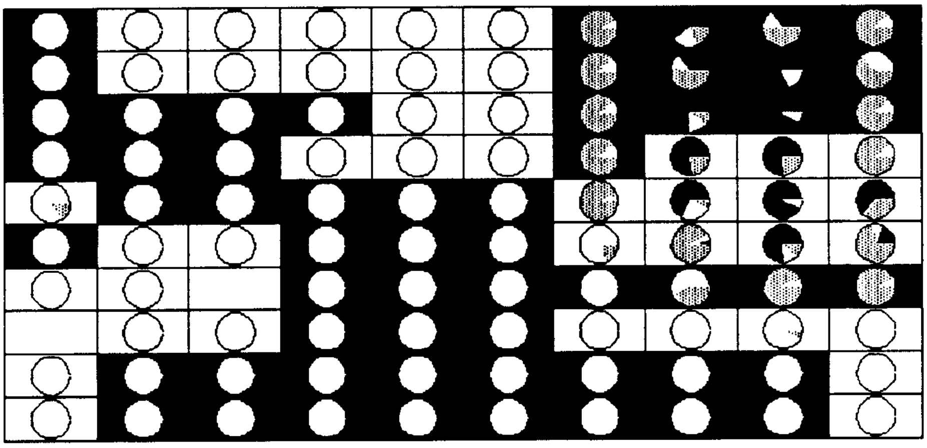

Fig. 6 : *In each cell the frequency pie of the variable DAY is represented. We can observe that units 8, 9, 18, 19, 28, 29, 38, 39, 48, 49, 50, 59 are mainly sundays. Units 7, 17, 27, 37, 47, 58, 59, 10, 20, 30, 40, 60, 70, are mainly saturdays. The other units days are exclusively devoted to the week-days.*

5 Analysis restricted to a small number of classes

As the Kohonen map is used in a non linear context, it is not judicious to realize a global linear analysis, but we can do it locally, that is if it is restricted to a part of the data set. We speak of local analysis when we only consider some neighboring classes. We define the best fit in the least-squares sense of this subset by a plane: it is the plane determined by the two first principal axes, as defined by a Principal Component Analysis.

As we are supposed to realize many local analyses and to do it quickly, it is important to choose a low computation time consuming method and we use here the EM algorithme to build the plane . See for example [10].

Then various illuminating graphical representations are at disposal, by considering various projections in this plane. Let be *P*(*A*) the plane which best fits a subset *A* of the observations.

1) *Projections of the observations of A in the plane P(A).*

We can point out the projections with any particular ploting symbols. We can use the name, or any other code which contains some information about the observations. It can be the number of the class, the corresponding modality of one of the qualitative variable, the quality of the projection computed through the cosine, etc.

In this way, it is easy to explain the dispersion on the plane and the relations between one unit and its neighbors (if the subset A contains more than one unit).

See in Fig. 7 and Fig. 8 two different ways of representing the projections. The subset A is the class number 23, the qualitative variable taken into account in Fig. 7 is the month and in Fig. 8, each point is coded by the square of its cosine with the plane.

2) *Projections of the points of others neighboring classes on the plane P(A).*

The projection of other neighboring classes on this plane can be very interesting. In Fig. 9, the neighboring classes 13, 22, 24, 33 of the unit 23 are projected in the plane defined by the unit 23, and we can have some idea about the mutual disposition of these classes. In Fig. 10, we can observe the repartition of the qualitative variable MONTH through the same five neighboring classes, centered around the class 23.

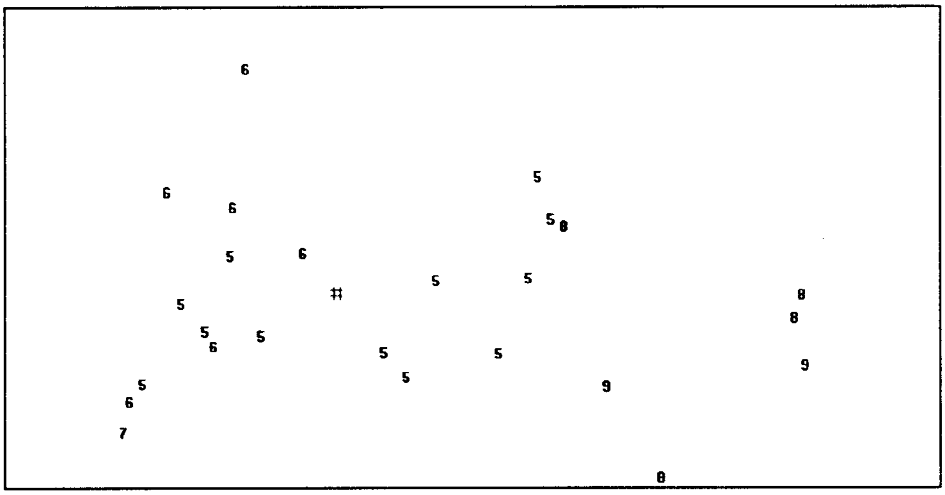

Fig. 7: *The points are coded by their month:*
May and June are at the left, August at the right

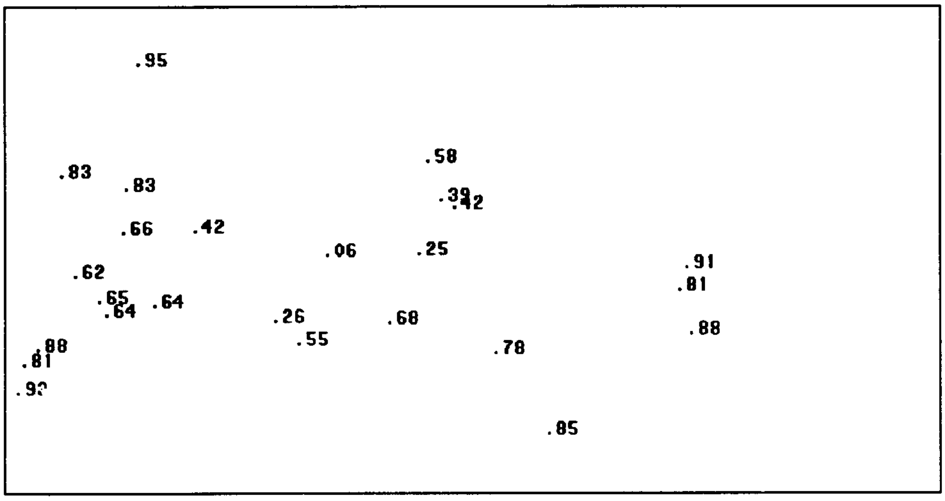

Fig. 8 : *The points are coded by the square of their cosine.*

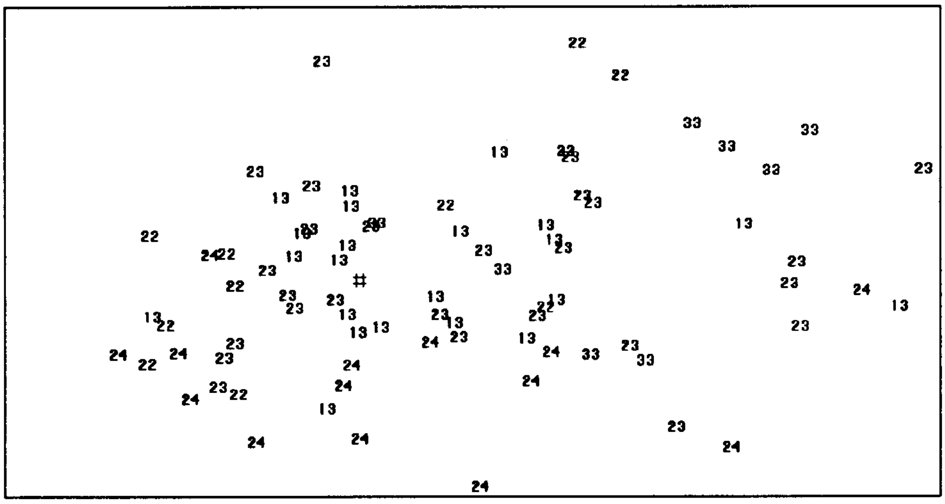

Fig 9 : *The elements of classes 23 and its four neighbors are ptojected in the plane defined by the class 23. The class 33 is located at the right corner, more or less separated from the other classes, as we could guess from Fig. 2.*

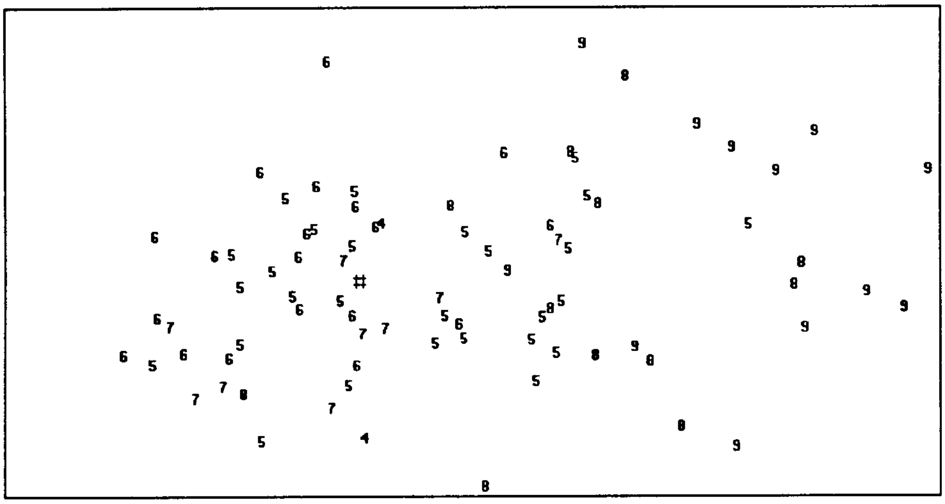

Fig. 10 : *The points are the same as in Fig.9, but they are coded by the MONTH. The results are coherent with those of Fig. 9 and Fig. 7. September characterizes class 33, May and June are at the left, August at the right, July is closer to May and June than to August and September.*

6 Conclusion and perspectives

We propose a general methodology to analyse multidimensional data, when a linear model is not satisfactory and when the observations are described by quantitative and qualitative variables. The present paper does not close the problem, since many other techniques can be combined in order to improve the analysis and the interpretation. In particular, we usually realize some multiple correspondence analysis [6] by adding to the qualitative variables a new one which represents the Kohonen classification or the macro-classification. In this way, we can propose a typology of the classes easy to interpret. We are now working to complete our software by incorporating many computations of standart statistics and in this way give not only visual tools but also quantitative ones.

Acknowledgements : We thank Corinne Muller (EDF, direction des Etudes et de la Recherche, Clamart, France) for fruitful discussions.

References

[1] T.W.Anderson, *An introduction to multivariate statistical analysis,* Wiley, 2nd ed., New York, 1984.

[2] F.Blayo, P.Demartines, Data analysis : how to compare Kohonen neural networks to other techniques ?, In *Proceedings of IWANN'91*, Springer, p. 469-476, 1991.

[3] M.Cottrell, J.C.Fort, Etude d'un algorithme d'auto-organisation, *Annales de l'Institut Poincaré*, Vol. 23, 1, 1-20, 1987.

[4] M.Cottrell, P.Letremy, E.Roy, E., Analysing a Contingency Table with Kohonen Maps: a Factorial Correspondence Analysis, *Proceedings of IWANN'93*, Springer Verlag, p. 305-311, 1993.

[5] M.Cottrell, J.C.Fort, G.Pagès, Two or three things that we know about the Kohonen algorithm, in *Proc of ESANN'94*, M. Verleysen Ed., D Facto, Bruxelles, p.235-244, 1995.

[6] M.Cottrell, S. Ibbou, Multiple Correspondence Analysis of a crosstabulations matrix using the Kohonen algorithm, in *Proc of ESANN'95*, M. Verleysen ED, D Facto, Bruxelles, p. 27-32, 1995.

[7] M.Cottrell, B.Girard, Y. Girard, C.Muller and P.Rousset, Daily Electrical Power Curves : Classification and Forecasting Using a Kohonen Map, *From Natural to Artificial Neural Computation, Proc. IWANN'95*, Springer, p. 1107-1113,1995.

[8] M.Cottrell, B.Girard, Y.Girard, M.Mangeas, Neural Modeling for Time Series: A Statistical Stepwise Method for Weight Elimination, *IEEE Tr. on Neural Networks,* Nov. 1995, Vol. 6, No 6, p. 1355-1364, 1995.

[9] M.Cottrell, E.de Bodt, A Kohonen Map Representations to Avoid Misleading Interpretations, in *Proc of ESANN'96*, M. Verleysen Ed., D Facto, Bruxelles, p.103-110, 1996.

[10] A.Dempster, N.Laird, D.Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. Roy Stat. Soc.,* B39, p. 1-38, 1977.

[11] A.Eydoux, D.Wuhl, Traitement statistique et économétrique des données ANPE sur le chômage et les activités réduites, Technical Report, Université Paris 1, 1996.

[12] F.Gardes, P.Gaubert, P.Rousset, Cellulage de données d'enquêtes de consommation par une méthode neuronale, *Preprint SAMOS #69, 1997.*

[13] P.Gaubert, S.Ibbou, C.Tutin, Housing market segmentation and price mechanisms in the Parisian metropolis, *International Journal of Urban and Regional Research,* to appear, 1995.

[14] T.Kohonen,. *Self-organization and Associative Memory*, 3°ed., Springer, 1993.

[15] T.Kohonen, *Self-Organizing Maps*, Springer Series in Information Sciences Vol 30, Springer, 1995.

Prépublications du SAMOS (depuis 1995)

1995

40 - Claude BOUZITAT, Gilles PAGES
Pour quelques images de plus..., 12 p.

41 - Jean-Claude FORT, Gilles PAGES
About the Kohonen algorithm: Strong or Weak Self-organisation? 15 p.

42 - Marie COTTRELL, Patrick LETREMY
Classification et analyse des correspondances au moyen de l'algorithme de Kohonen: application à l'étude de données socio-économiques. 10 p.

43 - Joël CHADOEUF, Xavier GUYON, Jian-Feng YAO
Sur l'ergodicité de l'estimation par Restauration-Estimation de modèles incomplètement observés. 11 p.

44 - Jean-Gabriel ATTALI, Gilles PAGES
Approximation of functions by perceptrons, a new approach. 11 p.

45 - Marie COTTRELL, Bernard GIRARD, Yvonne GIRARD, Corinne MULLER, Patrick ROUSSET
Daily electrical power curves : classification and forecasting using a Kohonen map. 8 p.

46 - Fabienne COMTE, Cécile HARDOUIN
Regression on log-regularized periodogram for fractional models at low frequencies. 19 p.

47 - Fabienne COMTE, Cécile HARDOUIN
Regression on log-regularized periodogram under assumption of bounded spectral densities: the non fractional and the fractional cases. 14 p.

48 - Patrick ROUSSET
Prévision des courbes demi-horaires au moyen d'une classification de Kohonen. 25 p.

49 - Smail IBBOU et Marie COTTRELL
Multiple Correspondence analysis of a crosstabulations matrix using the Kohonen algorithm. 6 p.

50 - Philippe JOLIVALDT
Schémas de discrétisation pour la simulation et l'estimation d'un CAR(2): une étude expérimentale. 22 p.

51 - Philippe JOLIVALDT
Utilisation de méthodes implicites pour la simulation et l'estimation de modèles CAR(2) . 14 p.

1996

52 - Samuel BAYOMOG
Estimation of a Markov field dynamic. 14 p.

53 - Morgan MANGEAS et Jian-feng YAO
Sur l'estimateur des moindres carrés des modèles auto-régressifs fonctionnels. 19 p.

54 - Marie COTTRELL, Florence PIAT, Jean-Pierre ROSPARS
A Stochastic Model for Interconnected Neurons. 17 p.

55 - Marie COTTRELL, Jean-Claude FORT, Gilles PAGES
Two or three mathematical things about the Kohonen algorithm. 31 p.

56 - Marie COTTRELL, Bernard GIRARD, Patrick ROUSSET

Forecasting of curves using a Kohonen classification. 14 p.

57 - Jean-Claude FORT, Gilles PAGES
Quantization vs Organization in the Kohonen S.O.M. 5 p.

58 - Eric de BODT, Marie COTTRELL, Michel LEVASSEUR
Réseaux de neurones en finance.33 p.

59 - Marie COTTRELL, Eric de BODT, Emmanuel HENRION, Ismaïl IBBOU, Annick WOLFS, Charles Van WYMEERSCH
Comprendre la décision à l'aide d'une carte de Kohonen. Une étude empirique. 16 p.

60 - Marie COTTRELL, Eric de BODT
Understanding the leasing decision with the help of a Kohonen map. An empirical study of the Belgian market. 5p.

61 - Marie COTTRELL, Eric de BODT, Philippe GREGOIRE
The relation between interest rate shocks and the initial rate structure: an empirical study using a Kohonen map. 16p.

62 - Marie COTTRELL, Eric de BODT, Philippe GREGOIRE
A kohonen map representation to avoid misleading interpretation. 8p.

63 - Marie COTTRELL, Eric de BODT
Analyzing shocks on the interest rate structure with Kohonen map. 6p.

64 - Marie COTTRELL, Eric de BODT, Philippe GREGOIRE
Simulating interest rate structure evolution on a long term horizon. A kohonen map application. 5p.

65 - Fabienne COMTE, Cécile HARDOUIN
Log-regularized periodogram regression. 22p.

66 - Jian Feng YAO
Simulation et optimisation sous contrainte par une dynamique de Metropolis. 6p.

67 - Morgan MANGEAS, Jian-feng YAO
On least squares estimation for nonlinear autoregressive processes. 16p.

68 - Carlo GAETAN, Xavier GUYON
Simulation des modèles de Gibbs spatiaux par chaine de Markov. 28p.

69 - François GARDES, Patrice GAUBERT, Patrick ROUSSET
Cellulage de données d'enquêtes de consommation par une méthode neuronale. 41p.

70 - Serge IOVLEFF, José R. León
High-Frequency approximation for the Helmholtz equation :a probabilistic approach. 14p.

1997

71 - Xavier Guyon et Jian-feng Yao
On description of wrong parametrisation sets of a model . 23p.

72 - Sandie Souchet
Schéma de discrétisation anticipatif et estimation du paramètre d'une diffusion. 36p.

73 - M. Cottrell, J.C. Fort, G. Pagès
Theoretic aspects of the SOM algorithm 22p.

74 - M. Benaïm, J.C. Fort & G. Pagès
Convergence of the one-dimensional Kohonen algorithm. 23p.

75 - C. Bouton & G. Pagès
About the multidimensional competitive learning vector quantization algorithm with constant gain. 36p.

76 - M. Cottrell, P. Rousset
The Kohonen algorithm: a powerful tool for analyzing and reprenting multidimensional quantitative and qualitative data 12p.