Analyzing and representing multidimensional quantitative and qualitative data: Demographic study of the Rhône valley. The domestic consumption of the Canadian families

Marie Cottrell, Patrice Gaubert, Patrick Letremy, Patrick Rousset

UMR 8595

Cahiers de la MSE

Analyzing and representing multidimensional quantitative and qualitative data: Demographic study of the Rhône valley. The domestic consumption of the Canadian families

Marie COTTRELL
Patrice GAUBERT
Patrick LETREMY
Patrick ROUSSET

1999.09

Maison des Sciences Économiques, 106-112 boulevard de L'Hôpital, 75647 Paris Cedex 13
ISSN : 1624-0340

Analyzing and representing multidimensional quantitative and qualitative data : Demographic study of the Rhône valley. The domestic consumption of the Canadian families.

Marie Cottrell, Patrice Gaubert, Patrick Letremy, Patrick Rousset

SAMOS-MATISSE, Université Paris 1
90, rue de Tolbiac, F-75634 Paris Cedex 13, France

1. INTRODUCTION

The SOM algorithm is now extensively used for data mining, representation of multidimensional data and analysis of relations between variables([1], [2], [5], [9], [11], [12], [13], [15], [16], [17]). With respect to any other classification method, the main characteristic of the SOM 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 SOM network. This feature allows to consider the resulting classification as a good starting point for further developments as shown in what follows.

But in fact its capabilities have not been fully exploited so far. In this chapter, we present some of the techniques that can be derived from the SOM algorithm: the representation of the classes contents, the visualization of the distances between classes, a rapid and robust two-level classification based on the quantitative variables, the computation of clustering indicators, the crossing of the classification with some qualitative variables to interpret the classification and give prominence to the most important explanatory variables. See in [3], [4], [8], [9] precise definitions of all these techniques. We also define two original algorithms (KORRESP and KACM) to analyze the relations between qualitative variables.

The paper is organized as follows : in sections 2 and 3, we present the main tools, in section 4 and 5, we show real applications in socio-economic fields.

2. THE MAIN TECHNIQUES

Let us give some notations : we consider a set of N observations, where each individual is described by P quantitative real valued variables and K qualitative variables. The main tool is a Kohonen network, generally a two-dimensional grid with n units, but the method can be used with any topological organization of the 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 examine here the delicate problem of the learning of the code vectors ([9], [16], [17]) which is supposed to be successfully realized from the N observations restricted to their P quantitative variables.

Classification:

After convergence, each observation is classified by a nearest neighbor method, (in R^P): observation l 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.

Representation of the contents:

The classes are represented according to the chosen topology of the network, along a chain, or on a grid, and all the elements can be drawn inside their classes. So it is possible to see how the observations are modified from a class to its neighbors and to appreciate the homogeneity of the classes.

Distances between classes

To highlight the true inter-classes distances, following the method proposed in [6], we represent each unit by an octagon inside the cell of the SOM map. The bigger it is, the closer it is to the border, the nearer the code vector is of its neighbors. This avoids misleading interpretations and gives an idea of the discrimination of the classes.

Two-level classification:

A hierarchical clustering of the n code vectors puts together the most similar SOM classes and provides a second classification into a smaller number of classes. These macro-classes create connected areas in the initial map, so the neighborhood relations between them are kept. This grouping together facilitates the interpretation of the contents of the classes.

Crossing with qualitative variable:

To interpret the classes according to an explicative qualitative variable, it is valuable to study the discrete distribution of its modalities in each class. We propose to draw inside each cell a frequency pie for example. So we make clear the continuity of the classes as well as the cutoffs. We also can associate to a SOM class the more frequent modalities of a qualitative variable and in this manner give a better description of the classes.

3. ANALYSIS OF RELATIONS BETWEEN QUALITATIVE VARIABLES

Let us define here two original algorithms to analyze the relations between qualitative variables. The first one is defined only for two qualitative variables. It is called KORRESP and is analogous to the classical Correspondence Analysis. The second one is devoted to the analysis of any finite number of qualitative variables. It is called KACM and is similar to the Multiple Correspondence Analysis. See [3], [4] for previous related papers.

For both algorithms, we consider a sample of individuals and a number K of qualitative variables. Each variable $k = 1,2, ..., K$ has m_k possible modalities. For each individual, there is one and only one modality. If M is the total number of modalities, each individual is represented by a row M-vector with values in {0, 1}. There is only one 1 between the 1st component and the m_1-th one, only one 1 between the m_1+1-th component and the (m_1+$m_{2)}$-th one and so on.

In the general case, where $M > 2$, the data are summarized into a Burt Table which is a cross tabulation table. It is a $M \times M$ symmetric matrix and is composed of $K \times K$ blocks, such that the (k,l)-block B_{kl} (for $k \leq l$) is the $(m_k \times m_{ll})$ contingency table which crosses the variable

k and the variable l. The block B_{kk} is a diagonal matrix, whose diagonal entries are the numbers of individuals who have respectively chosen the modalities 1, 2, ..., m_k , for modality k. From now on, the Burt Table is denoted by B.

In the case M=2, we only need the contingency table T which crosses the two variables. In that case, we set p (resp. q) for m_1 (resp. m_2).

3.1 The KORRESP algorithm

Let $M = 2$. In the contingency table T, the first qualitative variable has p levels and corresponds to the rows. The second one has q levels and corresponds to the columns. The entry n_{ij} is the number of individuals categorized by the row i and the column j. From the contingency table, the matrix of relative frequencies ($f_{ij} = n_{ij} / (\Sigma_{ij}\, n_{ij})$) is computed.

Then the rows and the columns are normalized in order to have a sum equal to 1. The row profile $r(i)$, $1 \le i \le p$ is the discrete probability distribution of the second variable when the first variable has modality i and the column profile $c(j)$, $1 \le j \le q$ is the discrete probability distribution of the first variable when the second variable has modality j. The classical Correspondence Analysis is a simultaneous weighted Principal Component Analysis on the row profiles and on the column profiles. The distance is chosen to be the χ^2 distance. In the simultaneous representation, related modalities are projected into neighboring points.

To define the algorithm KORRESP, we build a new data matrix D: to each row profile $r(i)$, we associate the column profile $c(j(i))$ which maximizes the probability of j given i, and conversely, we associate to each column profile $c(j)$ the row profile $r(i(j))$ the most probable given j. The data matrix D is the $((p+q) \times (q+p))$-matrix whose first p rows are the vectors $(r(i),c(j(i)))$ and last q rows are the vectors $(r(i(j)),c(j))$. The SOM algorithm is processed on the rows of this data matrix D. Note that we randomly pick the inputs among alternatively the p first rows and the q last ones and that the winning unit is computed only on the q first components in the first case, on the p last ones in the second case according to the χ^2 distance. After convergence, each modality of both variables is classified into a Voronoï class. Related modalities are classified into the same class or into neighboring classes. This method give a very quick, efficient way to analyze the relations between two qualitative variables. See [3] for real-world applications.

3.2 The KACM Algorithm

When there are more than two qualitative variables, the above method no longer works. In that case, the data matrix is the Burt Table B. The rows are normalized, in order to have a sum equal to 1. At each step, we pick a normalized row at random according to the frequency of the corresponding modality. We define the winning unit according to the χ^2 distance and update the weight vectors as usual. After convergence, we get an organized classification of all the modalities, where related modalities belong to the same class or to neighboring classes. In that case also, the KACM method provides a very interesting alternative to classical Multiple Correspondence Analysis.

The main advantages of both KORRESP and KACM methods are their rapidity and their small computing time. While the classical methods have to use several representations with

decreasing information in each, ours provide only one map, that is rough but unique and permits a rapid and complete interpretation. See [3], [4] and [7] for the details and financial applications.

4. DEMOGRAPHIC STUDY OF THE RHÔNE VALLEY

The data come from the project ARCHEOMEDES, supported by the EU, in collaboration with the laboratory P.A.R.I.S (University Paris 1).

We consider 1783 communes in the Rhône valley, in the south of France. This valley is situated on the two banks of the river Rhône. It includes some big cities (Marseille, Avignon, Arles, ...), some small towns, many rural villages. A large part is situated in medium mountains, in very depopulated areas since the so-called drift from the land. At the same time, in the vicinity of the large or small towns, the communes have attracted a lot of people who are working in urban employment. The goal of this study is to understand the relations between the evolution of the population and the professional composition in each communes.

The data include two tables. The first one gives the numbers of seven population census (1936, 1954, 1962, 1968, 1975, 1982, 1990). These numbers are normalized by dividing by their sum, to keep the evolution and not the absolute values. The second one contains the current numbers of working population, distributed among six professional categories (farmers, craftsmen, managers, intermediate occupations, clerks, workers). In this second table, the data are transformed into percentages and will be compared with the χ^2 distance.

The first step consists in defining two classifications of the communes from the two types of data. We use a Kohonen one-dimensional network (a chain) to transform the quantitative variables into qualitative ordered characters. First we classify the communes into 5 classes from the census data, and then into 6 classes from the professional composition data.

The first classification into 5 classes is easily interpretable. The classes are arranged according to an evident order : there are the communes with strong increase (*aug_for*), with medium increase (*aug_moy),* with relative stability (*stable*), with medium decrease (*dim_moy*), with strong decrease (*dim_for*). See in fig. 4.1 the code vectors and in fig. 4.2 the contents of the 5 classes.

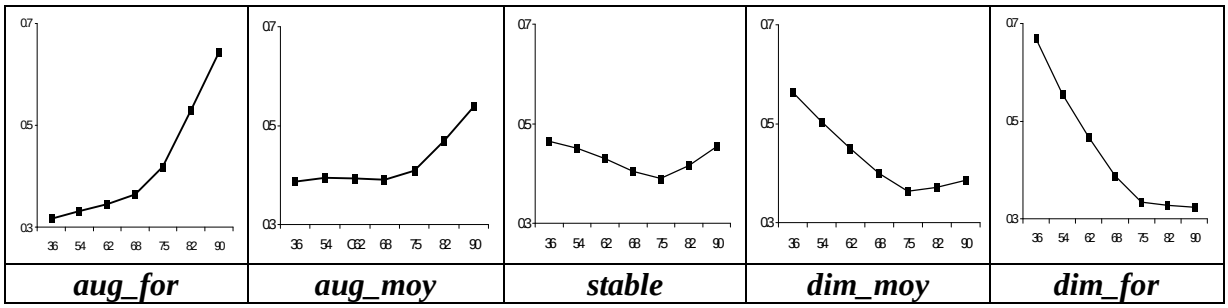

Fig. 4.1 : The code vectors of the 5 classes (first classification). The curves represent the population evolution along the seven census.

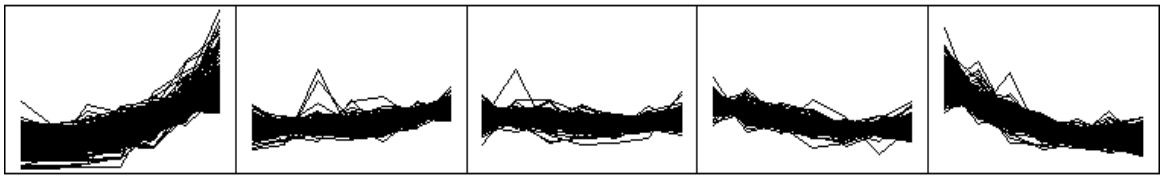

Fig. 4.2 : The contents of the 5 classes (first classification). In each class, all the communes vectors are drawn in a superposed way.

The 6 classes (A, B, C, D, E, F) provided by the second classification of the professional composition data, are a little more delicate to interpret. Actually they straightforwardly correspond to an order following the relative importance of the farmer category. Class A does not contain almost any farmer, while class F consists in communes with a majority of farmers (those are very small villages, but the use of the χ^2 distance restores their importance). See in fig. 4.3 the code vectors and in fig. 4.4 the contents of the 5 classes.

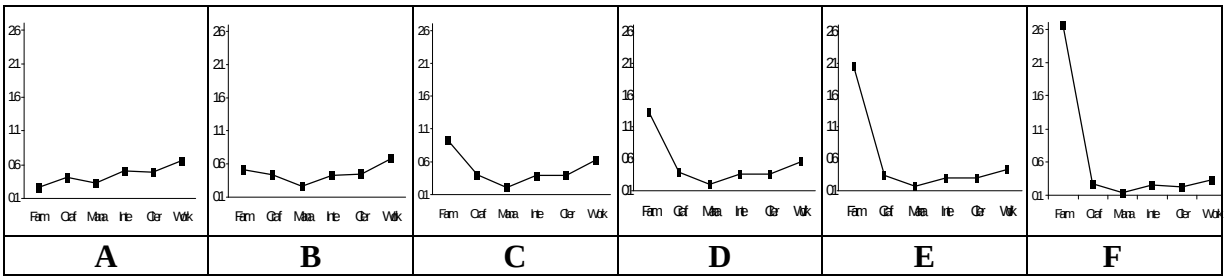

Fig. 4.3 : The code vectors of the 6 classes (second classification). The curves correspond to the (corrected) proportion of farmers, craftsmen, managers, intermediate occupations, clerks, workers.

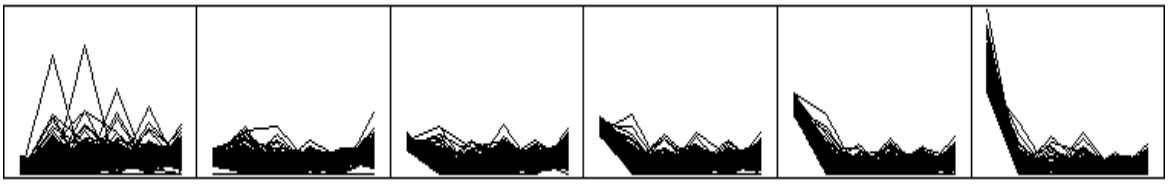

Fig. 4.4 : The contents of the 6 classes (second classification). Note that in class A, some communes are very specific, they do not have any farmer, but all their inhabitants belong to one or two categories.

From these two classifications, we compute the contingency table, see table 4.1. A quick glance shows a strong dependence between the row variables (census classes) and the column variables (professional classes).

Table 4.1 : Contingency Table which crosses the two classifications.

	A	B	C	D	E	F
aug_for	223	53	26	9	0	0
aug_moy	85	112	86	34	7	1
stable	80	100	112	113	54	22
dim_moy	29	50	55	80	56	57
dim_for	34	18	35	57	69	126

To analyze this dependence, we use a classical Correspondence Analysis and the KORRESP algorithm See in fig. 4.5 and table 4.3, the results.

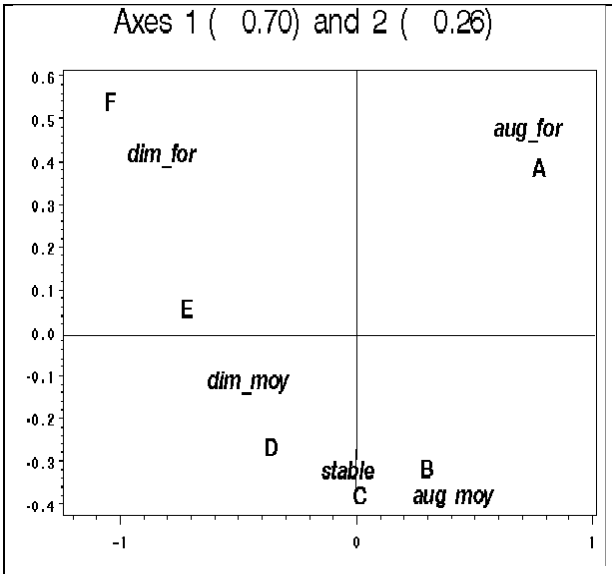

Fig.4.5 : The first projection of the modalities using a Factorial Correspondence Analysis

D	***stable***	**C**	
dim_moy			**B** ***aug_moy***
	E		
F ***dim_for***			**A** ***aug_for***

Table 4.3 : The two-dimensional Kohonen map with the results of KORRESP.

Both representations suggest to use a one-dimensional Kohonen network (a chain) to implement the KORRESP method. The results are shown in table 4.3.

Table 4.3 : The one-dimensional Kohonen map with the results of KORRESP.

A ***aug_for***	**B** ***aug_moy***	**C** ***stable***	**D** ***dim_moye***	**E**	**F** ***dim_for***

The conclusions are simple. The rural communes where the agriculture is dominant are depopulated, while the urban ones have an increasing population. The relations are very precise : we can note the pairs of modalities ((*dim_for*), F), ((*dim_moy*), D), ((*stable*), C), ((*aug_moy*), B), ((*aug_for*), A).

Actually, the SOM-inspired method is very quick and efficient, and gives the basic points of the information with only one representation. As to the classical correspondence method, it is also useful but its computation time is longer and it is usually necessary to examine several projections to have a complete analysis, since each axis represents only a percent of the total information.

5. The domestic consumption of the Canadian families

The data have been provided by Prof. Simon Langlois from the Université of Laval.

The purpose of the study is to define homogenous groups from the point of view of their consumption choices. The interest of such clustering is at least double. On the one hand, when one has successive surveys that include distinct individuals stemming from a same population, we can build a pseudo-panel, composed of synthetic individuals representative of the groups, which will be comparable from one survey to another.

On the other hand, it facilitates the matching of distinct surveys when each one provides different information about samples extracted from the same population. The constitution of groups which are homogenous for these data allows the linking of all the surveys. For example, it is possible to apply this method to match consumption surveys done for the same period with different samples (each sample is answered about the consumption of half-nomenclature). The matching is necessary to build complete consumption profiles. One has to notice that this method does not exactly correspond to the methodology proposed by Deaton, 1985, (13), who follows the same cohort from one survey to another by considering only individuals born at the same time.

When pseudo-panels are considered, one uses to build the clusters by crossing some significant variables. For example, to study the households consumption modes, the significant variables are the age cohort, the education level and the income distribution. Here, we use the Kohonen algorithm to define the clusters and apply it to the data of two consumption surveys. We also compare the results to those that we obtain from a standard classification.

5.1. The data

We consider two consumption surveys, performed by Statistiques Canada, in 1986 and 1990, with about 10 000 households which were not the same ones from one survey to the other. The consumption structure is known through a 20 functions nomenclature, described in Table 5.1.

Table 5.1 : Consumption nomenclature.

Alcohol; Food at home; Food away; House costs; Communication; Others; Gifts; Education ; Clothes; Housing; Leisure; Lotteries; Furniture; Health; Security; Personal Care ; Tobacco; Individual Transport; Collective Transport; Vehicles.

Each household is represented by its consumption structure, expressed as percentages of the total expenditure. The two surveys have been gathered in order to define classes including individuals which belong to one or the other year. So it will be possible to observe the dynamic evolution of the households groups which have similar consumption structures.

One can see that for any classification method, the classes contain in almost equal proportion data of the two surveys. It seems that there is no temporal effect on the groups, which simplifies the further analyses.

See in Fig. 5.1 the mean consumption structure for the 1992 survey.

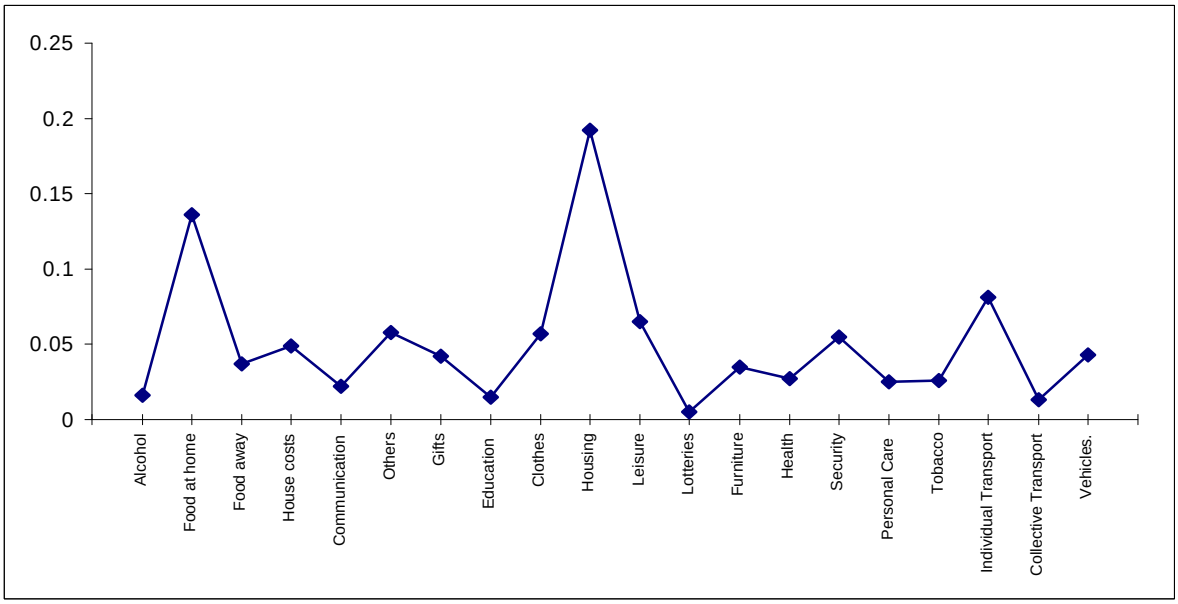

Fig. 5.1 Mean Consumption Profile in 1992.

5.2. The classes

We want to compare two clustering methods :

1) a SOM algorithm using a two-dimensional (8 × 8) grid, that defines 64 classes, whose number is then reduced to 10 macro-classes, by using a hierarchical clustering with the 64 code vectors, in order to get an easier interpretation of their contents.
2) a hierarchical classification into 10 classes with the Ward method.

5.3. The SOM classes and the macro classes

Fig. 5.2 represents the 64 SOM classes with their code vectors and the macro-classes which differ by their texture. First we note that, due to the topological conservation property of the SOM algorithm that the macro-classes group only neighboring SOM classes. In Fig 5.3, the distances between the SOM classes are drawn, following the method suggested in [6]. Observe that the grouping into 10 macro-classes respects the distance : the changes of macro-classes generally occur where the distances are larger.

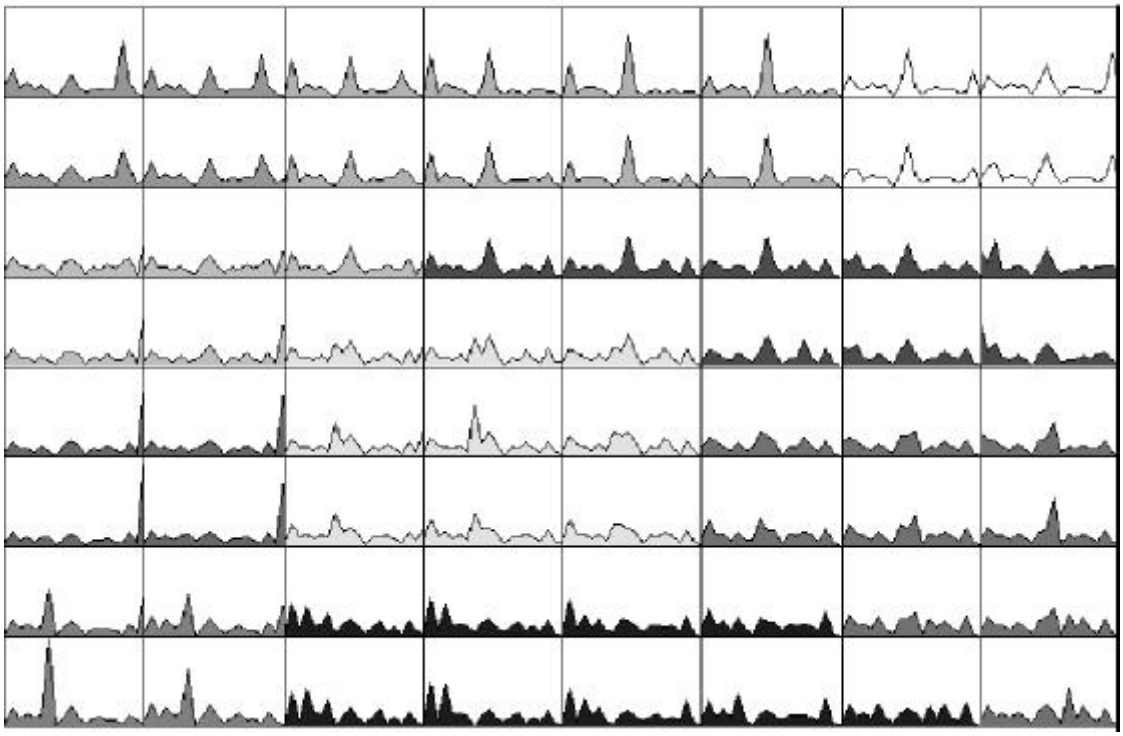

Fig. 5.2 : The 64 SOM classes, their code vectors and the macro-classes.

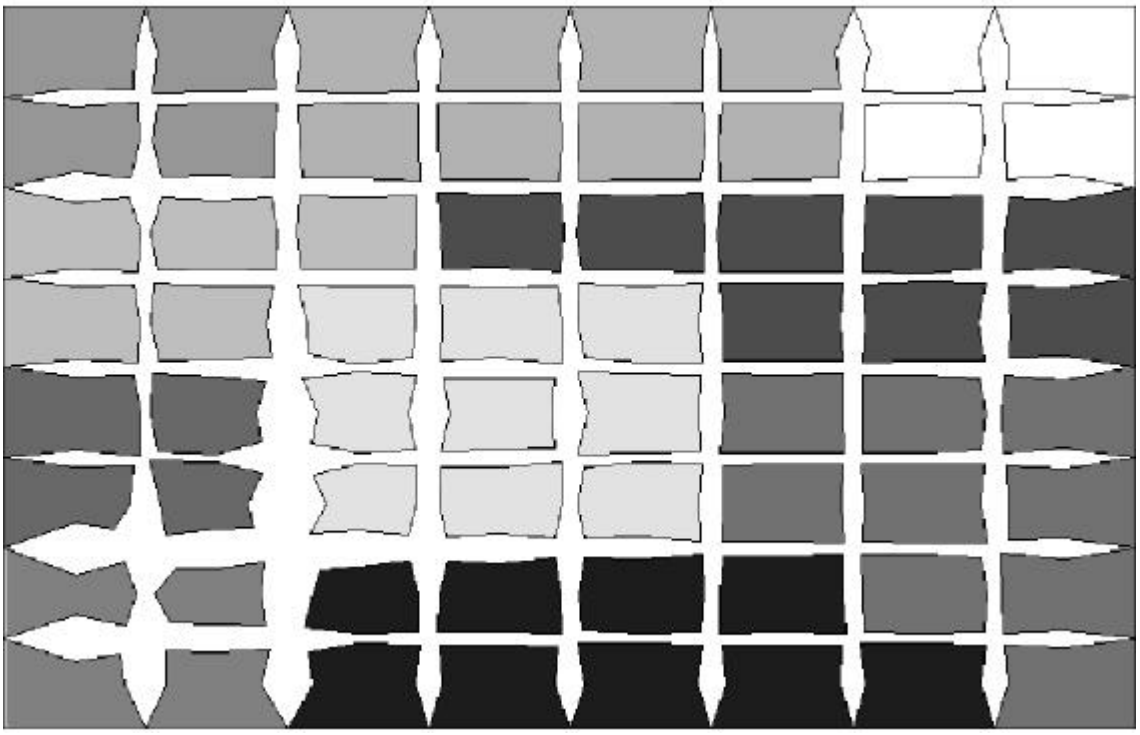

Fig. 5.3 : The distances between the 64 code vectors : in each direction, the classes are more distant when there is more white area.

The SOM classes could be analyzed, but it is difficult to keep and characterize 64 types of classes. Conversely, the 10 macro classes have well separated features. One can observe that :

1. the sizes of the macro-classes are about 600 to 700 households, or about 1200, except one with a little more than 400. This macro-class gathers only 4 SOM classes which have a very special profile (as it will be seen below).
2. in all the macro-classes, there are as many 1986 data as 1992 ones. So there is no significant effect of the year of the survey.
3. the mean profiles of the 10 macro-classes are well identified, and are different from the mean profile of the whole population.

Nine types of consumption items are at the origin of the differentiation of the macro-classes.

1. macro-class 5 is dominated by the *Housing* item (with a 38 %).
2. macro-class 9 is characterized by the importance of the *Housing* item (26 %) and the C*ollective Transport*.
3. for two macro-classes (1 and 2), the *Vehicle* purchase makes the difference. While the general mean value for this item is about 5 %, the value is 17 % in macro-class 1, and the other items are reduced in an homothetic way. In macro-class 2, the value is 36 %, and the housing expenditure is small, what corresponds to a large representation of the house-owners (71 % instead of 60 % in general).
4. the *Food Home* (20 %) and the *Others* items define the macro-class 7.
5. in macro-class 3, the *Security* (insurance) expenditure is the double mean value.
6. macro-class 10 corresponds to a large value of the *Gifts* item (25 %).
7. *leisure* item defines macro-class 8 (with 13 %), while *tobacco* defines macro-class 4 (with 12 %) and *education* is dominant in macro-class 6 (10 %).

The grouping into 10 macro-classes increases the contrast with respect to the mean consumption profile. All the SOM classes inside a macro-class have more or less the same features, with some specific characteristics.

5.4. Hierarchical clustering

If we consider the 10 classes defined by a hierarchical Ward clustering on the consumption profiles, the results are disappointing. The groups have unequal sizes, the differentiation between groups are more quantitative than qualitative, and poorer in information. For example, 4 groups have more than 1000 or 2000 elements, while the others have about 200 or 400. Among the 4 more important groups, 3 (the 1, 4, 6) have a mean profile similar to the general mean one, with only one component a little larger. Groups 2 and 7 correspond to a high *housing* expenditure, and cannot be clearly set apart, and so on.

Actually the correctly spotted groups are the small ones, while the others are not very different from one to another, and are similar to the general population. Furthermore, some specific behaviors, in particular those which are distinguished by a relatively large importance of the *Security* or *Education* expenditures, do not emerge in this clustering.

So from now on, we continue the analysis by using the SOM classification, followed by the grouping into 10 macro-classes.

5.5. Crossing with qualitative variables

To understand better the factors which determine the macro-classes, and to allow their identification, we use a graphic representation of some qualitative variables, that were not present in the classification.

For that, we use 4 qualitative variables, that give a socio-demographic description of the households :

1. the first one (*Wealth*) is a variable with 5 modalities (poor, quasi-poor, middle, quasi-rich, rich). This variable is defined by combining three simple criteria, the income distribution, the total expenditure, the food expenses, according to the age, the education level and the regional origin.
2. the second one (*Age*) is the age of the head of household, with 6 modalities (less than 30, 30-39, 40-49, 50-59, 60-69, more than 69).
3. the educational level (*Education*), with 5 levels (primary, secondary, post-secondary without diploma, post-secondary with diploma, university diploma).
4. the tenure status (*Tenure Status*) (owner or tenant).

For each SOM class, we compute the distribution of the four qualitative variables (which did not participate to the classification), and we represent it as a sector diagram (a pie) inside the cell. We observe that there is also a continuity for the variations of the socio-demographic variables distributions among the SOM classes. But they provide other information, different from the previous one.

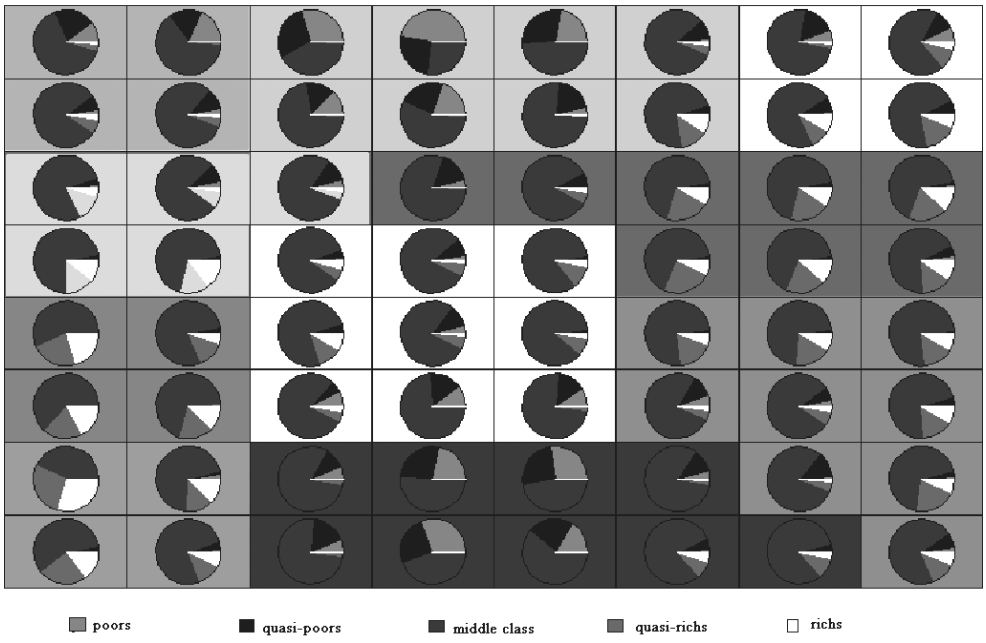

Fig. 5.4 : The distribution of the variable Wealth.

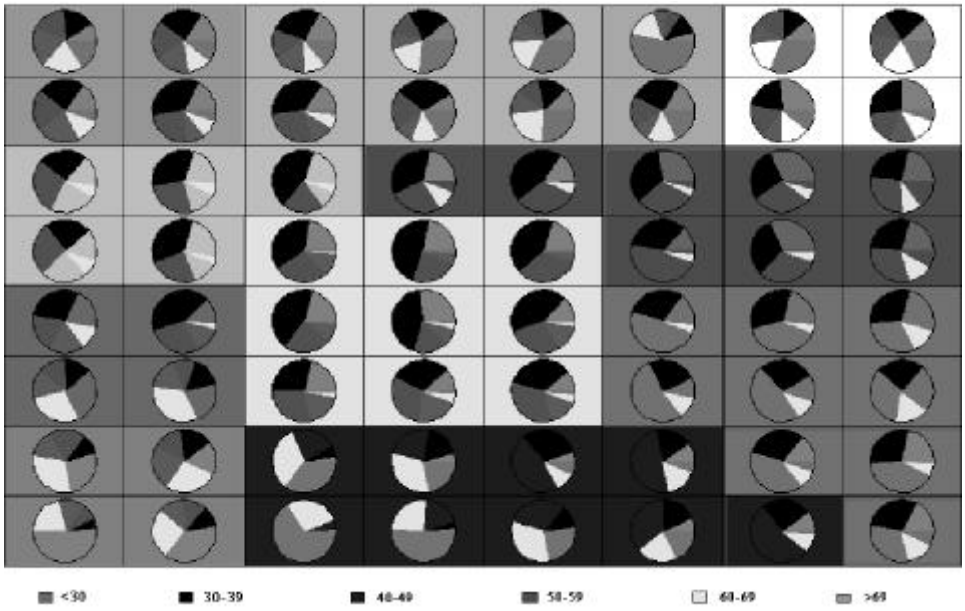

Fig 5.5 : The distribution of the variable Age

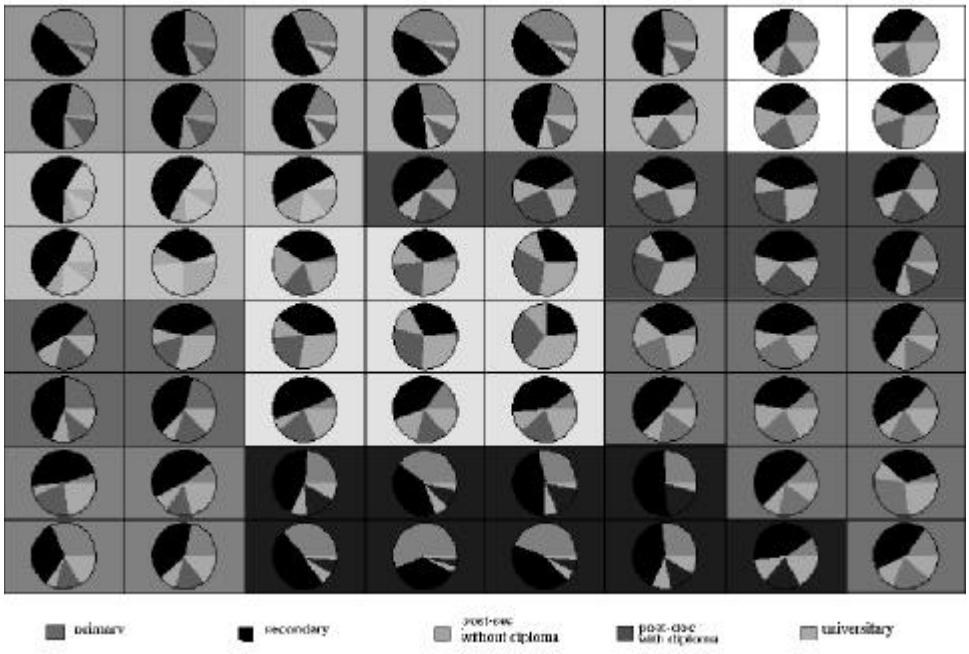

Fig. 5.6 : The distribution of the variable Education.

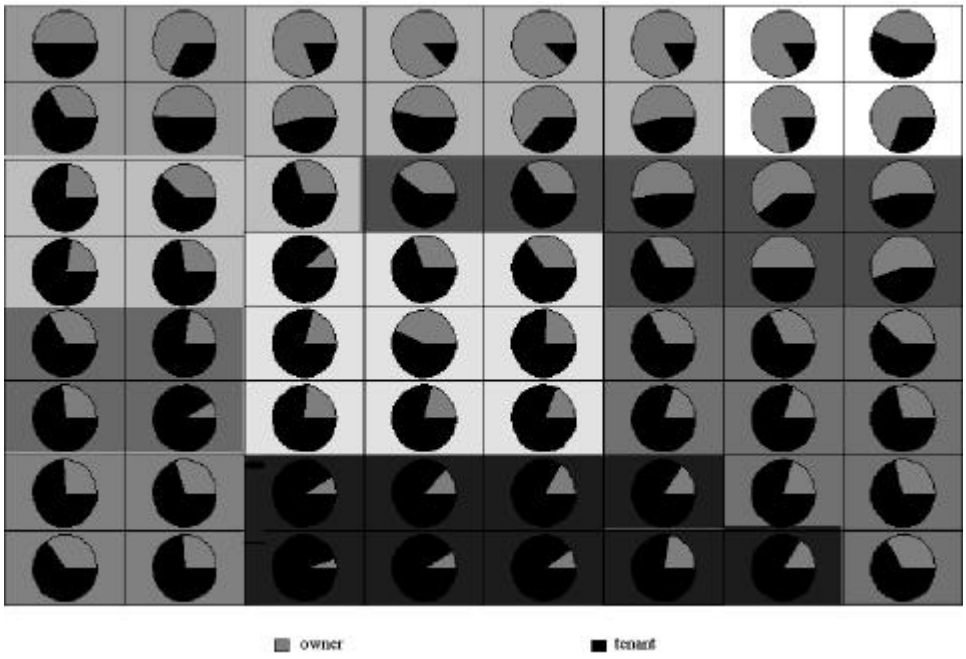

Fig 5.7 : The distribution of the variable Tenure Status

For example, the partitioning of the population according to the poverty-wealthy criterion (*Wealthy*), indicates that the classes having a strong proportion of rich or quasi-rich people are rather situated at the extremities of the diagonal right top - left bottom, the poor and quasi-poor being at the central area. At the same time, the opposition owner-tenant is distributed according to a simple opposition on this diagonal. The first ones are at the left bottom, the second ones at the opposite. We rediscover a well-known situation of poor people, who can be as well owner as tenant of their lodgings. It is possible to analyze in this way the four graphic representations. Actually, it is the combination of these characteristics that we have to examine to interpret the zones of the grid, as gathered by the classification into 10 classes.

6. Conclusion

The SOM algorithm is therefore a powerful tool to analyze multidimensional data and to help to understand the underlying structure. We are now working about local representation of the contents of a class in relation with the neighboring classes, in order to give an interpretation to the significant and discriminate variables. There is no doubt that the related data mining techniques will have a large development in many scientific fields, where one deals with numerous, large dimensioned data.

References

[1] F.Blayo, P.Demartines : Data analysis : How to compare Kohonen neural networks to other techniques ? In Proceedings of IWANN'91, Ed. A.Prieto, Lecture Notes in Computer Science, Springer-Verlag, 469-476, 1991.

[2] F.Blayo, P.Demartines : Algorithme de Kohonen: application à l'analyse de données économiques. Bulletin des Schweizerischen Elektrotechnischen Vereins & des Verbandes Schweizerischer Elektrizitatswerke, 83, 5, 23-26, 1992.

[3] M.Cottrell, P.Letremy, E.Roy : Analyzing a contingency table with Kohonen maps : a Factorial Correspondence Analysis, Proc. IWANN'93, J.Cabestany, J.Mary, A.Prieto Eds., Lecture Notes in Computer Science, Springer-Verlag, 305-311, 1993.

[4] M.Cottrell, S.Ibbou : Multiple correspondence analysis of a crosstabulation matrix using the Kohonen algorithm, Proc. ESANN'95, M.Verleysen Ed., Editions D Facto, Bruxelles, 27-32, 1995.

[5] M.Cottrell, B.Girard, Y.Girard, C.Muller, P.Rousset : Daily Electrical Power Curves : Classification and Forecasting Using a Kohonen Map, From Natural to Artificial Neural Computation, Proc. IWANN'95, J.Mira, F.Sandoval eds., Lecture Notes in Computer Science, Vol.930, Springer, 1107-1113, 1995.

[6] M.Cottrell, E. de Bodt : A Kohonen Map Representation to Avoid Misleading Interpretations, Proc. ESANN'96, M.Verleysen Ed., Editions D Facto, Bruxelles, 103-110, 1996.

[7] M.Cottrell, E. de Bodt, E.F.Henrion : Understanding the Leasing Decision with the Help of a Kohonen Map. An Empirical Study of the Belgian Market, Proc. ICNN'96 International Conference}, Vol.4, 2027-2032, 1996.

[8] M.Cottrell, P.Rousset :, The Kohonen algorithm: A Powerful Tool for Analysing and Representing Multidimensional Quantitative and Qualitative Data, Proc. IWANN'97, 1997.

[9] M.Cottrell, J.C.Fort, G.Pagès : Theoretical aspects of the SOM Algorithm, WSOM'97, Helsinki 1997, Neurocomputing 21, 119-138, 1998.

[10] A.Deaton : Panel data from time series of cross-sections, Journal of Econometrics, 1985.

[11] G.Deboeck, T.Kohonen : Visal Explorations in Finance with Self-Organization Maps, Springer, 1998.

[12] P.Demartines : Organization measures and representations of Kohonen maps, In : J.Hérault (ed), First IFIP Working Group, 1992.

[13] P.Demartines, J.Hérault: Curvilinear component analysis: a self-organizing neural network for non linear mapping of data sets, IEEE Tr. On Neural Networks, 8, 148-154, 1997.

[14] F.Gardes, P.Gaubert, P.Rousset: Cellulage de données d'enquêtes de consommation par une méthode neuronale, Preprint SAMOS # 69, 1996.

[15] S.Kaski: Data Exploration Using Self-Organizing Maps, Acta Polytechnica Scandinavia, 82, 1997.

[16] T.Kohonen: Self-Organization and Associative Memory, (3rd edition 1989), Springer, Berlin, 1984.

[17] T.Kohonen: Self-Organizing Maps, Springer, Berlin, 1995.