Neural Classification and « traditional » data analysis: an application to households' living conditions

Sophie Ponthieux*, Marie Cottrell**
* INSEE, Division "Conditions de vie des ménages",
sophie.ponthieux@insee.fr
**SAMOS-MATISSE, CNRS UMR 8595, Université Paris 1
cottrell@univ-paris1.fr

Abstract: The description, classification and « measurement » of living conditions present many difficulties. A very important one comes from the qualitative nature of the data, and the large number of characteristics that may be taken into account. For this reason, it is difficult to obtain a description that could give an overall view of the arrangements between the modalities, and be usable to breakdown the observations into a reasonable number of classes. In this paper, we propose several examples of the use of neural network techniques, precisely the Kohonen algorithm, to classify a population of households according to their characteristics in terms of living conditions.

1 Introduction

Since the 1970s in United-Kingdom, more recently in France, poverty is analyzed both in terms of income and in terms of living conditions, with a multi-dimensional approach. Living conditions include a great number of domains. Dickes (1994) lists ten of them: dwelling, durable, food, clothing, financial resources, health, social relations, leisure, education and work. Not all the existing studies include this complete set of domains. The choice of including or not one of those domains may be based on two arguments: in the first one, the main hypothesis is that the subjects are rationales in their behavior, which leads to select only the domains where privations are assumed to decrease or disappear when the financial resources are increased (Mack & Lansley, 1984); the second one is based on the notion of « standard » (Townsend, 1989), and leads to consider any domain as soon as all the subjects are, at least potentially, involved, i.e. following Dickes (1994, p.184), « all the households, whatever their composition or their situation in the life cycle ».

The main difficulty is that we have to deal with a great quantity of information that is mainly qualitative. We propose here to focus on two main objectives, taking into account only a relatively small number of domains (accommodation – in two parts: convenience, problems -, environment, durable and deprivations):

- obtain a good description of the characteristics: how they are combined, *i.e*. what are the most frequent associations between modalities, and in turn how the different sets of modalities are organized together,
- obtain an operational grouping of the observations when using only their

characteristics in the domains of the living conditions.

We compare two techniques: « traditional » data analysis (multiple correspondence analysis and clustering), and neural classification (Kohonen algorithm). The methods and data are shortly presented in *section 2.* Then we present in *section 3* two descriptions of the arrangement between the characteristics, the first one resulting from a Multiple Correspondence Analysis (MCA), the second one obtained with the Kohonen algorithm[1]. In *section 4* we compare different classifications of the observations: first a simple « score », then several classifications, obtained successively from a hierarchical clustering, then from Kohonen classification.

Due to the small space allowed for the papers, most of the graphs and detailed tables of results are gathered in an Appendix available on request by E-mail.

2 Method and Data

We suppose that the reader is familiar with the Kohonen algorithm. See for example Kohonen (1984, 1993, 1995), Kaski (1997), Cottrell, Rousset (1997) for an introduction to the algorithm and to its applications to data analysis. For our work, the main property of the Kohonen algorithm is the so-called topology conservation property. After convergence, in the resulting classification, similar data are grouped into the same class or into neighbor classes. This feature allows to represent the proximity between data, as in a projection, along the Kohonen map. As a further treatment, the Kohonen classes can be clustered into a reduced number of macro-classes (which only contain neighbor Kohonen classes) by using a classical hierarchical classification.

The data source used for this paper is the French part of the European Community Households Panel, here in its third wave (year 1996). It provides detailed information, both at the individual and the household level, about incomes (which allows to define an indicator of monetary poverty), and living conditions (material living conditions: dwelling, environment, durable goods, deprivations, but also financial living conditions: whether the monetary resources allow to live from "very comfortably" to "with great difficulty").

In what follows, the observations (households) are described according to their answers to questions covering the different domains of living conditions (*cf. infra*) and by a set of general characteristics: type of household, average age of the adults (persons aged 17 years and over), number of children under 17 years, type and location of the dwelling, financial living conditions, score for the material living conditions, indicator of monetary poverty. Only the observations with no missing variable for all theses descriptors are kept for the analysis, that is 6458 households. Only the variables describing the material living conditions are used in the classifications; we have kept all the information available[2]. The other variables are

[1] The SAS programs used for Kohonen classifications are due to Patrick LETREMY (SAMOS/MATISSE, Université de Paris 1)

2 This is rather different from what Dickes (1994) recommends ; according to this author, the choice of the

used to compare the different classifications. Finally, living conditions are described by 10 items about dwelling (5 about convenience and 5 about problems), 4 items about environmental topics, 6 items for the durable and 6 items about deprivations, a total of 26 dummy variables[3], that is to say 52 modalities.

This information is not always used under the same form: when we classify the characteristics, the inputs are a response table; when we classify the observations, the input is either the partial scores (score by domain) or the coordinates (obtained after a MCA, Multiple Correspondence Analysis) of the observations.

3. Description of the characteristics: classifications of the modalities

The modalities ended by 0 indicate the absence of problem (the households has a bath or a shower, has a separate kitchen, and so on). The suffix 1 corresponds to a problem (nor bath neither shower, no separate kitchen, and so on).

3.1. Results from a multiple correspondence analysis (MCA)

In all the representations (graph 1), most of the « positive » modalities appear in a very tight location. On the "negative" side, a group of modalities appears systematically isolated. It corresponds to what could be called « absence of a minimum set of conveniences in the dwelling » and combines no running hot water, no bath or shower, no indoor toilet. The other modalities are organized in several groups, often associating problems in the dwelling, bad quality of the environment, and scarcity of durable goods. It is interesting to notice that the « modern » durable (micro-wave, VCR) and the « traditional » ones (telephone, TV) appear to form two different sub-groups. It seems also that some deprivations (holidays and furniture) may have a particular status.

Finally, the MCA suggests more or less 4 groups of modalities:

- a first one grouping all the positive modalities (all the conveniences, no problem in the dwelling or in the environment, all the durable, no deprivation) with maybe one exception in the case of holidays and furniture;

- a second one where are associated absence of problems in the dwelling, bad quality of the environment, no holidays, and no possibility to replace worn-out furniture;

- and a third – possibly also a fourth group -, around the « absence of a minimum set of conveniences in the dwelling », added to no telephone and no TV set, combined (when representing the axis 3 and 4) with food deprivation (cannot buy meat/chicken/fish every second day if wanted).

---

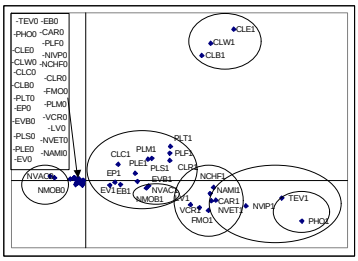
items must be based on a consensus about their necessity. This criteria is very often reduced to their frequency… because there is not a lot of information allowing to control for a consensus.

3 In the case of the durable, the answer distinguishes between "not having by choice" and "not having because cannot afford". Here, we have grouped « not having by choice » and « having », under the assumption that in this case there is no deprivation.

3.2 Results from a Kohonen classification

The modalities are classified first on a 10 x 10 grid (graph 2), then along a five classes string (graph 3). At first glance, it appears that the « positive » modalities are grouped at the top and on the right of the grid, and the « negative » ones at the bottom and on the left. The class in the bottom left cell corresponds to the « absence of a minimum set of conveniences in the dwelling » already mentioned with the MCA. It is here interesting to notice that the "very" negative modalities have no immediate neighbors, and are located on the grid quite on the exact opposite from their "positive" appearance. If we group the classes according to their closeness using a hierarchical classification, we obtain 3 sets of modalities, which are consistent with those resulting from the MCA. The classification along a string (one-dimensional Kohonen map) gives a very synthetic view of the associations between the modalities, and confirms the particular status of the holidays and replacement of furniture. A study of the profiles suggests a neat gradation of the negative modalities, which is an interesting result in that it could be usable for a "weighting" of the items.

Graph 1



Graph 2

	EB1 EP1		LV0		FMO0 VCR0		CAR0		CLB0 CLW0 CLE0
EVB1		EV1				NCHF0 NVIP0		PLT0 PHO0 TEV0	
				NAMI0	NVET0		EVB0		CLC0 PLE0 PLS0
CLB1 PLF1		PLM1				EB0 EP0		CLR0	
				NMOB0	NVAC0				PLF0 PLM0
PLS1		PLE1				EV0			
				VCR1			NMOB1		NVAC1
PLT1		CLC1				LV1			
					FMO1				NAMI1 NVET1
CLB1 CLW1 CLE1		PHO1 TEV1		CAR1		NCHF1		NVIP1	

Graph 3

CLB1	CLC1	EB1	EV0	CLB0	EVB0
CLW1	CLR1	EP1	NMOB0	CLC0	FMO0
CLE1	PLF1	EV1	NVAC0	CLR0	VCR0
PLT1	PLM1	NMOB1		CLW0	PHO0
FMO1	PLE1	NVAC1		CLE0	TEV0
PHO1	PLS1			PLF0	CAR0
TEV1	EVB1			PLM0	LV0
CAR1	VCR1			PLE0	NAMI0
NCHF1	LV1			PLS0	NCHF0
NVIP1	NAMI1			PLT0	NVET0
	NVET1			EB0	NVIP0
				EP0	

4. « Measuring » living conditions: classifications of the observations[4]

The simplest way to classify the observations according to their living conditions

[4] Table 1 provides a summary of the results : it gives the proportion of the super-classes obtained with each classification, and the concentration indicators (for an interpretation in terms of over-representation) of the different characteristics. We have added some general characteristics (not used for the classifications) in order to have a broader view of the population in the classes (type of dwelling, location, type of household).

consists in calculating a score of « bad points » for the whole set of items. The problem is there to determine the threshold to be used. This question is discussed in Lollivier & Verger (1997); they propose to use the rate of monetary poverty to define, by comparing this rate and the cumulative frequencies of the score, a threshold for living conditions "poverty". One problem with this method is that it does not allow to compare across the time the evolutions of these two measures of poverty, because, by construction, the two rates will be very close. Another problem is that it "cuts" the population into one group having "good" living conditions and one group having "bad" living conditions only on the criteria of one "bad point", whatever the item considered; therefore, even though it would be difficult to weight the items, it is clear that they are not all at the same level of "seriousness". So we have tried in what follows to obtain classes defined using the qualitative dimension of the information and independently from any exogenous threshold. The results are summarized in Table 1.

4.1 Classifications using the partial scores

The information used in this part is the scores obtained by each observation for each of the 5 domains.

4.1.1 Hierarchical classification

Ten clusters are then grouped in 3 super-classes; it gives about 15 % of the observations as living in the least favorable conditions (class 1), about 25 % in less unfavorable conditions (class 2), and 60 % in better living conditions (class 3). This classification appears to be consistent with the poverty rates (in monetary terms) in the three classes: about 30 % in class 1, 14 % in class 2, and only 1,6 % in class 3 (to be compared with 10,7 % for the whole sample). The three classes are differentiated also in terms of financial living conditions, neatly better when going from class 1 to class 3.

4.1.2 Kohonen classification

Here the Kohonen algorithm is utilized to obtain 10 classes, in turn grouped in 3 super-classes. Given the type of input (scores by domain), we have chosen to classify the observations along a string, in order to illustrate the *continuum*. The "progression" along the string appears neatly from best to worse living conditions. The proportions obtained in the three super-classes is slightly different from that obtained with the hierarchical classification: about 11 % of the observations appear as living in the least favorable conditions (class 3), again 11 % in less unfavorable conditions (class 2), and 78 % in better living conditions (class 1). In terms of monetary poverty, the classes are less different than in the previous classification. An interesting result is that some differences appear between classes 2 and 3: the scores in the domains of dwelling problems and bad quality of the environment are higher in class 2, while class 3 is characterized by higher scores in the domains of dwelling convenience, durable and deprivations.

4.2 Kohonen classifications using the observations coordinates (after a Multiple Correspondence Analysis)

The inputs are now the coordinates of the observations, resulting from a « traditional » MCA. So it corresponds to a transformation of the responses into quantitative values, (we use all the qualitative dimensions of the initial information). We use the Kohonen algorithm to classify the observations first on a grid, then along a string. Each one of these classifications is then grouped in 3 final super-classes.

4.2.1 Grid

One Kohonen classification is used here to build a 8 X 8 grid. The 64 classes obtained have then been grouped into 10, then 3 super-classes. The result is rather similar to that obtained in the classification of the modalities. The grouping into 10 and into 3 classes illustrates rather well the property of neighborhood that is one interest of this type of classification.

The final 3 super-classes give a distribution that is rather different from that obtained when using the partial scores: 15,2% (class C), 14,2% (class B) and 70,6% (class A). If class A appears clearly as the group having the « best » living conditions, the situations are not neatly different at first glance between classes B and C. This could be an interesting effect of taking into account the information at a very detailed level. Comparing classes B and C is particularly interesting: as for the partial scores, these two classes differentiate mainly in the domains of dwelling (in both terms of conveniences and problems) with higher scores in class C, while the scores are higher for durable and deprivations in class B. In terms of total score, poverty and financial living conditions, class B appears to be more disadvantaged than C.

So what we obtain here is a classification that suggests that the situations may be different in terms of living standards and in terms of living conditions, the difference coming mainly from the characteristics of the dwelling.

4.2.2 String (one dimensional Kohonen map)

Here, the observations are classified in 10 Kohonen classes along a string, then grouped in 3 super-classes.

Super-classes 1 to 3 represent respectively 70%, 14% and 16% of the observations; this is the classification that gives the higher proportion of unfavorable living conditions (if we add classes 2 and 3). It is also the classification that differentiates the least the classes in terms of monetary income, even though the poverty rate increases from class 1 to 3 (this indicating a greater heterogeneity of the incomes in class 3 which combines a higher income and a higher poverty rate than class 2). The global score increases from class 1 to 3, and between classes 2 and 3, the same differences appear for the domain of dwelling.

4.3. Comparison of the classifications

At first, and it is not very surprising, the proportion of households that can be said to

have unfavorable living conditions appear to be rather variable. Secondly, some neat differences appear according to the form of the information in input:

- classifications using as inputs the partial scores show generally a neat gradation in living conditions, ranking –by construction- from « worse » to « better »; but the Kohonen classification shows some differences between the domains that the hierarchical classification does not recreate.

- classifications based upon the observations coordinates show most interesting results: especially, the absence of minimum convenience in the dwelling distinguishes almost always one class, which is not systematically the most « underprivileged » in terms of global living conditions, poverty rate and financial living conditions. These classifications, because they use the qualitative dimension of the information, show some associations or specificities that do not appear if we « measure » living conditions with a « score ». It is particularly interesting in the case of dwelling convenience; an explanation could be that, given the high proportion of households having a minimum set of conveniences in the dwelling, the situation of not having this minimum set discriminates in itself a small proportion of households even though it were their only "negative" characteristic.

As for the general characteristics of the households (type of dwelling, location, type of household), the distributions appear rather consistent over all the classifications. Generally, unfavorable living conditions are more often than on average observed among households living in large structures, and among persons living alone and lone parents.

References

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

Cottrell, M., Fort, J.C. & Pagès, G. (1998) : Theoretical aspects of the SOM Algorithm, *Neurocomputing*, 21, p. 119-138.

Cottrell, M. & Rousset, P. (1997) : The Kohonen algorithm : a powerful tool for analysing and representing multidimensional quantitative et qualitative data, *Proc. IWANN'97*, Lanzarote.

Dickes, P (1994), *Ressources financières, bien-être subjectif et conditions d'existence*, *in* F.Bouchayer (ed.) Trajectoires sociales et inégalités, Eres.

Kaski, S. (1997) : Data Exploration Using Self-Organizing Maps, *Acta Polytechnica Scandinavia*, 82.

Kohonen, T. (1984, 1993) : *Self-organization and Associative Memory*, 3°ed., Springer.

Kohonen, T. (1995) : *Self-Organizing Maps*, Springer Series in Information Sciences Vol 30, Springer.

Lollivier, S & Verger, D (1997), *Pauvreté d'existence, monétaire ou subjective sont distinctes*, Economie & statistique n°308/309/310 « Mesurer la pauvreté aujourd'hui ».

Mack, J & Lansley, S (1984), *Poor Britain*, Allen & Unwin.

Mayer, S.E & Jencks, C (1989), *Poverty and the distribution of material hardship*, Journal of Human Resources vol.24.

Townsend, P (1989) : *Deprivation*, Journal of Social Policy, vol.16.

Table 1 – Summary of the results

		Scores		**Classes based on partial scores**						**Classes based on coordinates**					
				Hierarchical			**Kohonen**			**Grid+3 super-classes**			**String+3 super-classes**		
		0	1	1	2	3	1	2	3	A	B	C	1	2	3
Distribution (%)		89.2	10.2	60.7	24.6	14.8	77.6	10.9	11.5	70.6	14.2	15.2	70.0	13.8	16.1
Concentration indicators *(proportion –or mean- for a given class / proportion –or mean – for the whole sample)*															
Score for the material living conditions	Total (all domains)	0.7	**3.2**	0.5	**1.2**	**2.5**	0.5	**1.8**	**3.9**	0.3	1.0	**4.4**	0.7	**2.3**	**1.1**
	Dwelling, convenience	0.8	**2.9**	0.8	0.6	**2.5**	0.4	**4.3**	**2.0**	0.7	**1.5**	**1.7**	0.8	**1.2**	**1.5**
	Dwelling, problems	0.9	**2.2**	0.9	0.8	**1.7**	0.8	**2.2**	**1.2**	0.9	**1.4**	**1.1**	0.6	**1.1**	**2.6**
	Environment	0.7	**3.7**	0.2	**1.6**	**3.4**	0.7	**1.2**	**2.7**	0.6	**2.9**	**1.2**	0.7	**1.3**	**1.9**
	Durables	0.7	**3.2**	0.1	**1.7**	**3.4**	0.4	0.8	**5.6**	0.4	**3.4**	**1.7**	0.8	**1.3**	**1.8**
	Deprivations	0.7	**3.2**	0.4	**1.2**	**2.9**	0.7	**1.9**	**2.4**	0.6	**2.2**	**1.5**	0.7	**1.3**	**1.9**
Monthly income per c.u.*(a)*		1.0	0.6	**1.2**	0.8	0.6	**1.1**	0.8	0.6	**1.1**	0.7	0.9	1.0	0.8	0.9
Monetary poverty	poor*(b)*	0.7	**3.2**	0.4	**1.3**	**3.1**	0.6	**1.4**	**3.2**	0.6	**2.3**	**1.8**	0.9	**1.1**	**1.6**
	non poor	1.0	0.7	**1.1**	1.0	0.7	1.0	1.0	0.7	**1.1**	0.8	0.9	1.0	1.0	0.9
Financial living conditions	very difficult	0.5	**5.4**	0.1	0.9	**4.8**	0.5	**1.3**	**4.1**	0.4	**3.7**	**1.4**	0.7	0.7	**2.5**
	difficult	0.8	**2.3**	0.4	**1.5**	**2.4**	0.8	**1.4**	**2.0**	0.8	**2.1**	1.0	0.9	**1.3**	**1.4**
	rather difficult	1.0	**1.1**	0.7	**1.5**	**1.1**	0.9	**1.3**	**1.2**	0.9	**1.2**	**1.1**	0.9	**1.4**	1.0
	rather comfortable	**1.1**	0.2	**1.3**	0.7	0.2	**1.1**	0.8	0.4	**1.1**	0.4	0.9	**1.1**	0.8	0.7
	comfortable and very c.	**1.1**	0.1	**1.5**	0.4	0.1	**1.2**	0.6	0.2	**1.2**	0.2	0.8	**1.2**	0.4	0.8
Type of dwelling	House, isolated	1.0	0.6	**1.1**	0.9	0.7	**1.1**	0.6	0.7	**1.1**	0.7	0.9	**1.1**	0.9	0.7
	House, in a neighborhood	1.0	1.0	1.0	**1.1**	1.0	1.0	**1.1**	1.0	1.0	1.0	1.0	1.0	**1.1**	1.0
	Structure <10 unitss	0.9	**1.4**	0.9	1.0	**1.3**	0.9	**1.5**	**1.4**	0.9	**1.2**	**1.4**	0.9	**1.1**	**1.2**
	Structure >=10 units	1.0	**1.4**	0.9	1.0	**1.3**	0.9	**1.3**	**1.2**	0.9	**1.4**	1.0	0.9	1.0	**1.4**
	Other	1.0	**1.3**	1.0	**1.1**	**1.1**	0.9	1.1	**2.0**	0.8	0.7	**2.3**	1.0	**1.2**	0.8
Location	Rural town	1.0	0.9	1.0	**1.1**	1.0	1.0	0.9	**1.1**	1.0	0.9	**1.1**	**1.1**	1.0	0.6
	City <10000 inh	1.0	0.8	**1.1**	1.0	0.8	1.0	0.9	0.8	1.0	0.9	1.0	1.0	0.8	1.0
	10000 to <100000 inh	1.0	1.0	0.9	**1.1**	1.0	1.0	1.0	1.0	1.0	**1.1**	1.0	1.0	**1.1**	1.0
	100000 to <2000000 inhh	1.0	**1.2**	1.0	0.9	**1.1**	1.0	1.0	**1.1**	1.0	**1.1**	0.9	0.9	1.0	**1.3**
	Paris area	1.0	0.9	**1.1**	0.8	0.9	1.0	**1.3**	0.9	**1.1**	0.9	0.8	0.9	**1.1**	**1.3**
Type of household (children taken into account if <25 years old)	Person living alone	0.9	**1.5**	0.8	**1.2**	**1.4**	0.9	1.0	**1.6**	0.9	**1.2**	**1.4**	0.9	**1.3**	**1.1**
	Couple without child	**1.1**	0.5	**1.1**	0.9	0.6	**1.1**	0.8	0.6	**1.1**	0.7	0.9	1.0	0.8	0.9
	Couple with child(ren)	1.0	0.8	**1.1**	0.8	0.8	1.0	**1.1**	0.6	**1.1**	0.9	0.8	**1.1**	0.8	0.9
	Lone parent family	0.9	**1.9**	0.6	**1.3**	2.0	0.9	**1.1**	**1.9**	0.9	**1.9**	0.8	0.8	**1.3**	**1.5**
	Other	1.0	0.9	1.0	**1.1**	1.0	1.0	0.9	1.3	1.0	1.0	**1.1**	**1.1**	0.9	0.6

(a) Equivalence scale : 1 – 0.5 – 0.3

(b) Poverty threshold at 50 % of the median income per c.u.