Professional Trajectories of Workers using Disconnected Self-Organizing Maps

Etienne Côme[1], Marie Cottrell[2], and Patrice Gaubert[3]

[1] IFSTTAR - Bâtiment Descartes 2,
2, Rue de la Butte verte, 93166 Noisy le Grand Cedex, France
etienne.come@ifsttar.fr
[2] SAMM - Université Paris 1 Panthéon-Sorbonne
90, rue de Tolbiac, 75013 Paris, France
marie.cottrell@univ-paris1.fr
[3] ERUDITE, Université Paris 12,
61, avenue du Géneral De Gaulle, 94010 Créteil, France
patrice.gaubert@u-pec.fr

Abstract. Using the Panel Study of Income Dynamics (PSID) collected on the period 1984-2003, we study the situations of American workers with respect to employment. The data include all heads of household (men or women) as well as the partners who are on the labor market, working or not. They are extracted from the complete survey by computing a few relevant features which characterize the worker's situations. To perform this analysis, we suggest to use a Self-Organizing Map (Kohonen algorithm) with specific topology. In this paper we present a new topology for SOM based on a planar graph with disconnected components (called D-SOM) which is especially interesting for clustering. Each component takes the form of a string and corresponds to an organized cluster.
From this clustering, we study the dynamics at the individual level, that is the trajectories of the individuals among the classes during the observed period. Then we estimate the transition probability matrices for each studied year and the corresponding stationary distributions.
Finally, we try to give an answer to the question: is there a significant change in 1992 (new economic policies after the Reaganomics).
Keywords: Kohonen algorithm, planar graphs, labor market, Markov chains

1 Introduction

The aim of this study is to identify and to analyze the succession of situations occupied by workers on a modern labor market (1984-2001). The mainstream theory presents mechanisms to explain the level of labor furnished for a specified compensation, the stability of the relation between a firm and a worker and its evolution over time (a career). These mechanisms are not observed in the most real situations. To identify the diversity of situations in terms of activity is the first step of the study.

A situation is defined by quantitative variables:

- global quality of a job, full time job for the whole year, wages, seniority in the same job versus
- positions with more or less precarious conditions: wages lower than the average, part time jobs, jobs for short periods, on-call jobs, current practice of a second job

Working on individual data we construct a classification of situations observed every 2 years on a specific labor market during two consecutive periods of nine years. With the characteristics of a small set of major situations, it is possible to define the successive localizations of each individual for each studied year of the two periods. That is what we called trajectories between situations.

We need to study the temporal changes, during both sub-periods: 1984-1992 and 1993-2001. It must be possible to answer some important questions linked to the evolution of the macroeconomic environment: in 1992 the end of Reaganomics and the beginning of Clinton period which leads to a global reduction of unemployment. What is the impact of this reduction of unemployment and is there a significant change at the individual level?

This article follows another paper [2] but contains necessary material (and possibly redundant) to be self-contained. It is organized as follows : first, in Section 2, the data and the notations used throughout the paper are presented. The methodology and the global architecture of the proposed procedure are described in Section 3. Each step is defined and results on real data are given in Sections 4 to 7.

2 The Data: first period (1984, 86, 88, 90, 92) and second period (93, 95, 97, 99, 2001)

We use the PSID (Panel Study of Income Dynamics), dividing the observations in two sub-periods in order to solve the trade-off we meet: observe a number of workers large enough to obtain statistical indicators representative of the whole population, from one hand, and from the other hand, to keep only individuals present along the period to identify trajectories.

We create a sample for each period (1984-1992, 1993-2001) but, with the hypothesis that the main situations have the same characteristics in these periods, with differences in levels only, we make the classification with all the observations together.

In the PSID data, we select households for which the head (man or woman) is present every studied year of the period but separately for each sub-period. The administrative rule is that if there is a male in the household he is the head, if not the head is a woman. Fortunately quite the same variables concerning the activity on the labor market are available for the wife of the head, if there is one. Retrieving this information we constitute set of individuals (around 4 500 per year) observed every two years in each sub-periods, with a proportion of women close to the one observed in the whole population.

An observation consists of a couple (year, individual). It is described by 8 quantitative variables and 4 qualitative variables. See Table 1 for the list of variables and their meaning. j

Name	Description	Type
nbbhtrav	Number of worked hours per week	Quant
nbstrav	Number of worked weeks	Quant
nbschom	Number of unemployed weeks	Quant
nbsret	Number of weeks out of labor market	Quant
salhor	Wages per hour	Quant
nbex	Number of extra jobs	Quant
hortex	Number of hours worked in extra jobs	Quant
anctrav	Seniority in present work in monthes	Quant
sex	Sex	Qual
naiss	Year of birth	Qual
pro	Professional occupation	Qual
bri	Branch of industry	Qual

Table 1. Variable name, description and type.

3 Disconnected Self-Organizing Maps, D-SOM

Following Come et al. (2010) [2], we use a light variant of the classical SOM ([3], in order to get a map which is composed of several disconnected one-dimensional strings. Each string will contains data which are similar at a rough level and that are displayed in ordered disposition.

To get this topology, it is necessary to define a neighborhood structure which is different from the classical one. Graph theory allows us to define such structures as noted by several authors like [1, 4]. If the used graph can be represented in dimension 2, we will still have the advantages of Self-Organizing Maps for visualization and data mining.

See figure 1 an example of disconnected neighborhood structure that we define here.

This topology has a special interest: when the map consists of not connected parts, the "cooperation" step of the algorithm only concerns the units which belongs to the same component as the winning unit. The competition step is not modified, so that the algorithm complies a double goal :

1. to group the observations into macro-classes corresponding to the different connected components of the graph ;
2. to organize the units inside the macro-classes.

The code-vectors are denoted by m_{ij}, $i \in \{1, \ldots, K\}$, $j \in \{1, \ldots, n_i\}$, where K is the number of disconnected components and n_i is the size of component

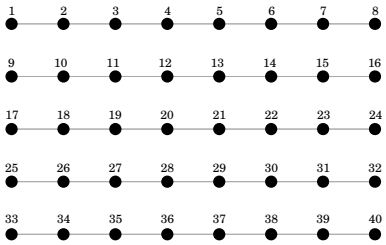

Fig. 1. Bi-dimensional representation of a disconnected map with 5 strings of 8 units.

i. Then the distance $d((i,j),(i',j'))$ between classes (i,j) and (i',j') is define as the shortest path distance in the graph. It is equal to $+\infty$ if $i \neq i'$. The code-vectors which do not belong to the same macro-class as the winning unit are not updated by the cooperation step.

The algorithm can be written as below:

1. The code-vectors are randomly initialized in the data space ;
2. at each step t, the code-vectors are updated $\mathbf{m}_{ij}(t)$ in the following way :
   - one observation $\mathbf{x}_{t+1}$ is randomly drawn and we achieve two steps;
   - *Competition,* the winning unit is computed for the l'observation $\mathbf{x}_{t+1}$ by:

$$[i^*(t+1), j^*(t+1)) = \arg \min_{i \in \{1,\dots,K\}, j \in \{1,\dots,n_i\}} ||\mathbf{x}_{t+1} - \mathbf{m}_{ij}(t)||; \quad (1)$$

   - *Cooperation,* the code-vectors of the winning unit and of its neighbors (which necessarily belong to the same macro-class (i^*, j^*)) are updated by:

$$\mathbf{m}_{i^*j}(t+1) = \mathbf{m}_{i^*j}(t) + \alpha(t) h(t, (i^*, j^*), (i^*, j)) \left[\mathbf{x}_{t+1} - \mathbf{m}_{i^*j}(t)\right], \quad (2)$$

   where t is the number of iteration, $\alpha(t)$ is the learning rate and $h(t, (i^*, j^*), (i^*, j))$ is the neighborhood function at step t between classes (i^*, j^*) and (i^*, j).

In conclusion, by imposing a limitation of the cooperation which only acts inside the macro-classes and by keeping a competition between all units, this algorithm allows us to get a classification into a given number of macro-classes which are themselves self-organized.

There exists other methods to get well-separated classes, see [5] for example. But our approach is different since we do not look for building an adjacency matrix between the code-vectors by repeating many runs of the SOM algorithm. Contrarily, we impose an a priori adjacency matrix which defines non-connected classes.

This kind of topology is well adapted in the frame of the labor market segmentation, since one looks for a segmentation into macro-classes well discriminated, easy to describe, split into organized classes. In a general case, the question of

the choice of the number of macro-classes is guided by a priori argument if there exists theoretical reasons. In our case we chose 5 macro-classes which is the best choice to get contrasting and well identified situations.

Let us now describe the results that we get using this topology for our data.

4 The map, description of the clusters

Figure 2 shows the about 45000 couples (year, individual) represented by a 8-vector, classified into 5 disconnected macro-classes, themselves composed of 8 units.

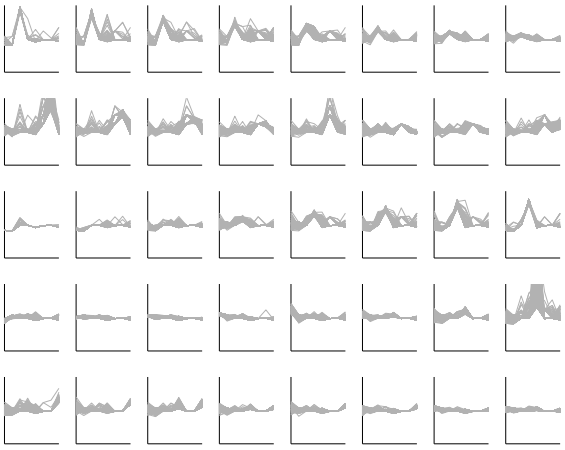

Fig. 2. D-SOM map with 5 macro-classes of 8 units

By computing the arithmetic means (see Table 2 and Table 3) of the eight variables used to make the classification, it is easy to emphasize the contrasts between the macro-classes, (the five strings):

- macro-class 1: precarious, part-time employment and unemployment
- macro-class 2: people having extra jobs (one or more) to obtain a sufficient standard of living
- macro-class 3: people most of the time out of the labor market (discouraged, ill, or for family reasons, and retired people in period 2)
- macro-class 4: full employment with very short seniority in the present place
- macro-class 5: full employment with the highest compensation and seniority (about 18 years)

	C1 (1493)	**C2** (2736)	**C3** (3756)	**C4** (7443)	**C5** (6686)
nbhtrav	36.99	40.69	8.28	**42.03**	41.63
nbstrav	27.71	47.15	5.68	**48.50**	47.29
salhor	8.25	11.87	2.62	12.80	**14.28**
nbschom	**22.13**	0.54	0.29	0.14	0.11
nbsret	0.96	0.19	**9.48**	0.10	0.01
anctrav	28.95	82.43	5.58	30.20	**168.14**
nbex	0.05	**1.13**	0.01	0.00	0.00
hortex	8.58	**384.83**	1.35	0.77	0.12

Table 2. Mean values for each variables by macro-class, period 1; the figures in bold are the maximum values for each variable, the figures between brackets are the class sizes.

	C1 (531)	**C2** (2171)	**C3** (6108)	**C4** (7099)	**C5** (7081)
nbhtrav	35.76	41.07	4.36	**42.25**	41.59
nbstrav	31.11	47.32	3.49	**48.13**	47.31
salhor	14.61	19.70	1.74	19.44	**20.25**
nbschom	**21.16**	0.34	0.06	0.10	0.06
nbsret	2.21	0.60	**3.27**	0.26	0.25
anctrav	29.94	108.62	3.50	29.36	**214.26**
nbex	0.04	**1.11**	0.00	0.00	0.00
hortex	7.99	**397.37**	0.31	0.61	0.00

Table 3. Mean values for each variables by macro-class, period 2; the figures in bold are the maximum values for each variable, the figures between brackets are the class sizes.

These findings are obtained for the two sub-periods.

Figure 3 contains five subplots which present the evolution of the code-vectors along a macro-class from unit one to unit eight. All the variables are centered and reduced and are drawn on the same scale $[-5, 10]$. This representation confirms our description.

Figure 4 presents the 8 variables on the whole D-SOM map, with 5 macro-classes of 8 units each one.

5 Transitions

The study of trajectories followed by individuals observed over a period of nine years is obtained computing the transition matrix: it shows the probability to be in one class at year $t + 2$, starting from another class at year t. See Table 4 for the first period and Table 5 for the second one.

The most evident result is that the major part of a class has not moved between year t and year $t + 2$: the important exception to this rule is class

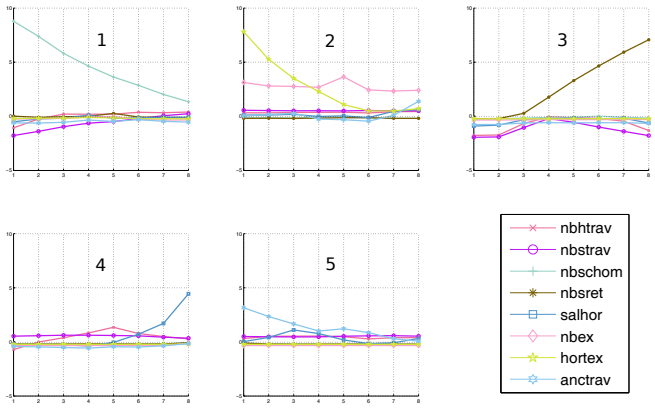

Fig. 3. Multivariate profiles of the different macro-classes.

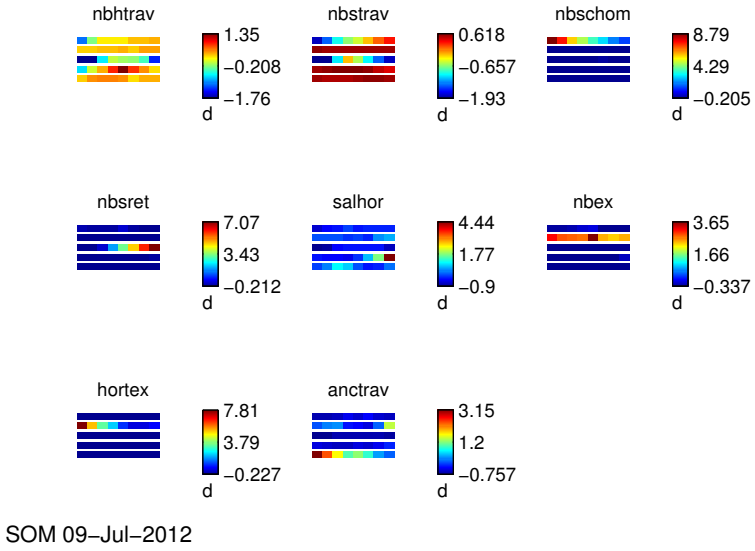

Fig. 4. 8 variables on the whole D-SOM map, with 5 macro-classes of 8 units each one

one, in each sub-period. A large part of the precarious, unemployed, part-time workers have changed to good jobs two years later, and this phenomenon is even more important in second period.

Of course, the most stable class over both sub-periods is class 5, the one with very stable jobs. The great proportion observed in period 2 of people staying in class 3 is probably due to the effect of people aging while they are observed and definitely leaving the labor market. It is interesting to notice that in the second period a proportion significantly smaller of the class 1 is staying in this class, that is the worst situation. It must be the effect the growing flexibility introduced in the US economy.

	C1	**C2**	**C3**	**C4**	**C5**
C1	0.26	0.10	0.14	**0.40**	0.10
C2	0.03	**0.52**	0.03	0.26	0.17
C3	0.06	0.03	**0.66**	0.22	0.02
C4	0.06	0.08	0.07	**0.63**	0.16
C5	0.03	0.06	0.02	0.09	**0.79**

Table 4. Transition matrix, period 1, values in bold are maxima.

	C1	**C2**	**C3**	**C4**	**C5**
C1	0.09	0.09	0.16	**0.59**	0.08
C2	0.02	**0.45**	0.04	0.31	0.19
C3	0.01	0.01	**0.85**	0.11	0.01
C4	0.02	0.08	0.08	**0.66**	0.16
C5	0.02	0.05	0.03	0.15	**0.75**

Table 5. Transition matrix, period 2, values in bold are maxima.

6 Limit and empirical distributions

For each period, we can compare the observed distributions of individuals across the five macro-classes to the theoretical limit distributions, computed under the hypothesis that everything in the environment stays unchanged. The limit distribution is estimated by iterating the transition matrix, which converges, as shown by Markov Chain Theory, to a matrix whose all rows are the same. So that the transition probabilities do not depend anymore on the starting value. We see Table 6 that there is a change between period 1 and 2. The theoretical and

observed distributions are closer, one to the other, in period 2 than in period 1. This indicates that the system has become more stable, i.e. the successive distributions are approximately the same during period 2.

	C1	**C2**	**C3**	**C4**	**C5**
Empirical distribution for the first period	0.07	0.12	0.17	0.34	0.30
Limit distribution for the first period	0.06	0.12	0.13	0.31	0.38
Empirical distribution for the second period	0.02	0.09	0.27	0.31	0.31
Limit distribution for the second period	0.02	0.08	0.29	0.33	0.28

Table 6. Empirical and limit distributions, period 1 and 2.

7 Some results by gender

Here is the distribution of men and women on the map. See Figure 5

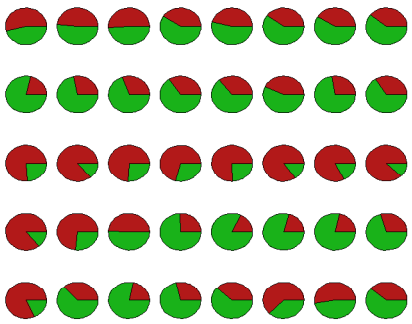

Fig. 5. Men in light grey (green) and women in dark grey (red), explanations are in the text

Knowing that men and women are in close proportions in the two samples (like they are in the whole population), it is easy to observe that men are more numerous than the average in the five last units of macro-class 4 and in the most part of macro-class 5 (macro-classes 4 and 5 correspond to the best situations).

They are also the main part of macro-class 2, the one where workers have one extra job or more.

At the same time women are a great proportion, from 2/3 to 4/5, of those who, for a while or definitely, are out of the market (macro-class 3).

If we look at the transitions matrix for the 2 periods and the genders, one can see that (we do not display them for lack of space):
- the major fact for both genders is the withdrawal from the market in the second period (people discouraged and/or not registered as present on the labor market or retired)
- men are leaving part-time jobs or true unemployment (macro-class 1) to obtain full-time unstable jobs (macro-class 4) in a greater proportion in second period, women move in a greater proportion towards the class 3, as in period 1
- from macro-class 4 of full-time jobs without seniority, women are more leaving towards the withdrawal (macro-class 3), while this move is very weak for men.

8 Conclusion

From this real-world example, we showed that using a Disconnected Self-Organized Map algorithm facilitates the clustering of numerous data, by providing a segmentation into easy-to-interpret clusters, themselves being divided into well-organized classes. Then the classification can be interpreted at two levels that are of interest. The number of macro-classes is a priori defined, equal to the number of clusters that the experts have identified. This is the "supervised" part of the algorithm. The interest of SOM which is a non-supervised method is that each cluster can be described by the population it contains, and that we can retrieve its main characteristics.

At the second level, each cluster is, in turn, split into micro-classes, which are mainly organized according to the value of one of the input variables. This fact provides a refined description of each cluster's population. In the future, we want to study an automatic method to select the topology, the number of macro-classes, the size of the strings, in the framework of model selection.

References

1. Barsi, A.: Neural self-organization using graphs. In: Proceedings of the Third International Conference Machine Learning and Data Mining in Pattern Recognition, Leipzing (Germany). Lecture Notes in Artificial Intelligence, vol. 2734, pp. 343–352. Springer (July 2003)
2. Côme, E., Cottrell, M., Verleysen, M., Lacaille, J.: Self organizing star (sos) for health monitoring. In: Proceedings of the European conference on artificial neural networks, Bruges (Belgium). pp. 1341–1346 (April 2010)
3. Kohonen, T.: Self-Organizing Maps. Information sciences, Springer (1995)
4. Pakkanen, J., Iivarinen, J., Oja, E.: The elvoving tree - analysis and applications. IEEE Transactions on Neurals Networks 17, 591–603 (2006)
5. Resta, M.: Assessing the efficiency of health care providers: a som perspective. In: Advances in Self-Organizing Maps. Lecture Notes in Computer Science, vol. 6731, pp. 30–39 (June 2011)