Career-path analysis using drifting Markov models (DMM) and self-organizing maps

S. Massoni[1] , **M. Olteanu**[2], **P. Rousset**[3]

[1] CES, Université Paris 1
112 Bd de l'Hopital, Paris, France
[2] SAMM, Université Paris 1
90 rue de Tolbiac, Paris, France
[3] CEREQ, 10 Place de la Joliette, Marseille, France

Abstract

Analyzing school-to-work transitions is an important challenge for the specialists of the labor-market. The aim of this paper is to study the insertion of graduates and to identify the main career-paths typologies. We introduce a new methodology for clustering career-paths by combining statistical estimation of non-homogeneous Markov chains with self-organizing maps. The proposed methodology is tested on real-life data issued from the survey "Generation 98" elaborated by CEREQ, France (http://www.cereq.fr/)

Keywords: Career paths, categorical data, drifting Markov model, self organizing maps.

1 Introduction

Analyzing school-to-work transitions is an important issue for the specialists of the labor-market. Identifying the main typologies, characterizing the transitions from one status to another, measuring the impact of the fixed-termed contracts on the obtaining of a permanent contract, assessing the role of public temporary contracts in the labor-market insertion are all challenging topics that the economists are attempting to answer. All these questions become even more important today, when, in the current global context, youth insertion in the labor market is delicate (for example, 22.3% of the young people under 25 were unemployed during the first semester of 2009 in France).

The aim of this paper is to study the insertion of graduates and to identify the main career-paths typologies. Let us recall that a career-path is defined as a sequence of labor-market statuses, recorded monthly. The data used for this study was coded into three statuses: stable employment (permanent contracts), temporary employment (fixed-termed, apprenticeship, temporary public and on-call contracts) and out-of-the-labor-market (unemployed, inactive or still in the education system). The military service status was deleted from the career-paths, since it brings no information on the professional insertion.

The question of identifying career-path typologies has already been studied in the literature. The-state-of-the-art is dominated by two approaches that seem to be currently used. The first approach consists in performing a factorial analysis in order to transform the categorical variables corresponding to the statuses into quantitative variables and then apply usual clustering algorithms for Euclidean data ([5]). The second approach consists in computing adapted distances for the data ([9], [2]) and then cluster it using an hierarchical tree. Both approaches have some drawbacks: in the first case, the use of factorial analysis implies quite strong hypothesis on the linearity of the data, while in the second case, hierarchical clustering is not suited for large data sets and does not provide any tools for displaying and visualizing the results. Moreover, since a career-path is a categorical time-series, dependence between the variables should be taken into account when computing the distance. Here again, two approaches seem to dominate the literature: on the one hand, compute the distance between two paths using optimal matching ([1]) and, on the other hand, use first-order Markov chains to model the career-paths ([4]). Both methods can be criticized. For the first, transition, insertion and deletion costs have to be defined a priori by the user. The definition of these costs is subject to lively debates in the scientific community. For the second, the use of homogeneous Markov chains is not entirely appropriate since it seems natural that the probabilities of transition are not the same at different moments of the trajectory and that the time spent in each status has an important influence on the transitions. Let us also note the recent approach proposed by ([6]) where a new distance taking into account the time-series structure of the data is introduced.

In this paper, we introduce a new methodology for clustering career-paths. The idea is to combine statistical estimation of non-homogeneous Markov chains with self-organizing maps. Thus, the non-homogeneous temporal dependence will be taken into account when clustering the data. The rest of the paper is organized as follows: Section 2 is devoted to a description of the proposed methodology, while Section 3 contains the results on the data "Génération 98" provided by CEREQ (France). The conclusion and some perspectives are presented in the last section.

2 Methodology

Let us now introduce the main steps of the methodology. In the first place, we shall recall that the data to be clustered are categorical time-series of different lengths. Instead of transforming the categorical variables into quantitative via factorial analysis, we will associate a non-homogeneous Markov chain to each career-path. Then, the transition probabilities will be estimated for every path and the estimates, which are quantitative, will be used as clustering variables. Next, we shall detail the estimation procedure for the Markov chains and we will briefly recall the self-organizing-maps algorithm.

2.1 Drifting Markov models with linear drift

As previously mentioned, a non-homogeneous Markov chain will be associated to each career-path. The estimates of the transition probabilities for every chain will characterize the corresponding career-path. This approach allows us to take into account the dependence between the variables in the paths and also the impact of time on the transitions between statuses.

In a previous study, we used a methodology combining optimal matching and self-organizing maps [8]. Although the results had been promising, we weren't quite satisfied with the dissimilarity computed by optimal matching since it assumed homogeneous transitions between states and ignored the impact of time. Since the transitions between the statuses in a career-path may be supposed linked to the moment of time at which the statuses occur, introducing non-homogeneous transition matrices appeared as a more reasonable and realistic assumption.

A first solution would have been to use hidden Markov models (HMM) in which different transition matrices are fitted on different homogeneous parts of the trajectory. However, this approach was not possible, since the lengths of the career-paths were too short and wouldn't have permitted to obtain reliable estimates. Instead of using HMM, drifting Markov models (DMM) were considered as an alternative.

DMM ([10]) represent a wide class of non-homogeneous Markov models characterized by an initial and a final transition matrix linked together by a linear or a polynomial deterministic function. Thus, the transition probabilities for any position in the sequence may be written either as a linear or as a polynomial function of the initial and final transition probabilities. The link function must be chosen a priori by the user, while the parameters to be estimated are the initial and the final transition probabilities. For this study, a linear function was selected. Next, we shall briefly recall the DMM model with linear drift and the estimation procedure.

Let $X = (X_t)_{t\in\{0,\ldots,n\}}$ be a sequence of random variables. A DMM of order 1 is defined by its transition matrix:

$$\Pi_{\frac{t}{n}} = \mathbb{P}(X_t = v | X_{t-1} = u)$$

with $u, v \in \mathcal{A}$, where $\mathcal{A}$ is the state space. In the case of a linear drift, the transition matrix $\Pi_{\frac{t}{n}}$ can be written as a linear combination of the initial and the final transition matrices, Π_0 and Π_1, as follows:

$$\Pi_{\frac{t}{n}} = \left(1 - \frac{t}{n}\right)\Pi_0 + \frac{t}{n}\Pi_1$$

The estimates for Π_0 and Π_1 are obtained using a point-by-point method which minimizes the sum of prediction errors :

$$\sum_{t=1}^{n}\sum_{u\in\mathcal{A}}\sum_{v\in\mathcal{A}} \mathbf{1}_{\{X_{t-1}=u\}}\left(\Pi_{\frac{t}{n}}(u,v) - \mathbf{1}_{\{X_t=v\}}\right)^2$$

The estimates of the transition probabilities in Π_0 and Π_1 are then computed by Lagrange minimization:

$$\hat{\Pi}_0(u,v) = \frac{B_2(u)C_1(u,v) - B_1(u)C_2(u,v)}{A_1(u)B_2(u) - A_2(u)B_1(u)}$$

$$\hat{\Pi}_1(u,v) = \frac{A_1(u)C_2(u,v) - A_2(u)C_1(u,v)}{A_1(u)B_2(u) - A_2(u)B_1(u)}$$

with

$$A_1(u) = 2\sum_{t=1}^{n}\mathbf{1}_u\left(1 - \frac{t}{n}\right)^2,\ A_2(u) = 2\sum_{t=1}^{n}\mathbf{1}_u\left(1 - \frac{t}{n}\right)\left(\frac{t}{n}\right),$$

$$B_1(u) = 2\sum_{t=1}^{n} \mathbf{1}_u \left(1 - \frac{t}{n}\right)\left(\frac{t}{n}\right),\ B_2(u) = 2\sum_{t=1}^{n} \mathbf{1}_u \left(\frac{t}{n}\right)^2,$$

$$C_1(u,v) = 2\sum_{t=1}^{n} \mathbf{1}_{uv} \left(1 - \frac{t}{n}\right),\ C_2(u,v) = 2\sum_{t=1}^{n} \mathbf{1}_{uv} \left(\frac{t}{n}\right).$$

2.2 Self-organizing maps with missing data

Once the transition matrices are estimated and every career-path has an assigned vector of transition probabilities, the data is clustered using a self-organizing map algorithm ([7]). The clustering is done on the unnormalized quantitative variables represented by the transition probabilities and using the Euclidean distance.

Let us recall that a Kohonen map carries out, at the same time, data clustering (vector quantization) and nonlinear projection. The input data is projected onto a two-dimensional grid and the projection respects the topology of the data: two vectors which are close in the initial space will be projected in the same class or in neighbor classes. Hence, clustering via self-organizing maps provides more information than other clustering methods, since it gives insights on the proximity of the clusters and a two-dimensional representation of the data.

There is still an issue to be considered before clustering with self-organizing maps: missing data. Indeed, if one of the statuses is missing from the career-path (if for example, after graduation, the subject obtains directly a permanent contract, the temporary-contract status will not be observed in his career-path) then, the transition probabilities associated to that status will not be estimated. This means that the Kohonen algorithm should handle the case where several values in an input vector are missing. We used the approach introduced by [3], where the winning prototype for each input is computed only on the available variables. The SAS programs developed by P. Letremy (http://samos.univ-paris1.fr/Programmes-bases-sur-l-algorithme) were used for performing this step of the analysis.

3 Real-life data - "Generation 98" survey

To test our methodology, we performed an analysis on the data issued from the survey “Generation 98” elaborated by CEREQ, France. The data set contains information on 16040 young people having graduated in 1998 and monitored during 94 months after having left school. The labor-market statuses have nine categories, labeled as follows: “permanent-labor contract”, “fixed-term contract”, “apprenticeship contract”, “public temporary-labor contract”, “on-call contract”, “unemployed”, “inactive”, “military service”, “education”. The following stylized facts may be highlighted by a first descriptive analysis of the data:

- permanent-labor contracts represent more than 20% of all statuses after one year and their ratio continues to increase until 50% after three years and almost 75% after seven years;
- the ratio of fixed-terms contracts is greater than 20% after one year on the labor market, but it is decreasing to 15% after three years and then seems to converge to 8%;

- almost 30% of the young graduates are unemployed after one year. This ratio is decreasing and becomes constant, 10%, after the fourth year.

Two transformations were made in the original data set:

- the "military service" status was suppressed since it brings no essential information on the employment or unemployment situation and since it may be considered as a parenthesis in the career.
- in order to obtain reliable estimates for the transition probabilities and because of the relatively short length of the career-paths (94 values before suppressing the military service), the number of categories was reduced to three: "stable employment", "short-time employment", and "out-of-the-labor-market" [1].

Then, the recoded data set can be represented as in Fig.1:

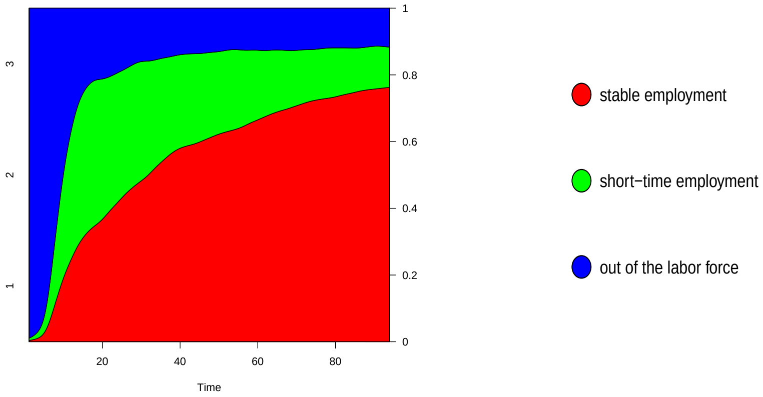


Figure 1: Labor-market structure

The proposed methodology was run in two steps. Firstly, the drift-Markov models corresponding to each career path were estimated. To each input vector correspond two estimated transition matrices (the initial and the final ones) of size three. The elements of the matrices were stored as vectors of size 18. Hence, the initial 16040x94 categorical data set was transformed into a 16040x18 quantitative data set. Secondly, a self-organizing maps modified to handle missing values was run on the transformed data set. The projection structure chosen for the map was a 10×10-rectangular grid.

The resulting map is summarized in Figure 2. For each class, we represented the evolution of the career-paths throughout the 94 months of study. Figure 3 contains the number of inputs in each cluster.

[1] These three categories are constructed as follows:

– *stable employment*: permanent-labor contract

– *short-time employment*: fixed-term contract, apprenticeship contract, public temporary-labor contract, on-call contract

– *out of the labor force*: unemployed, inactive, education

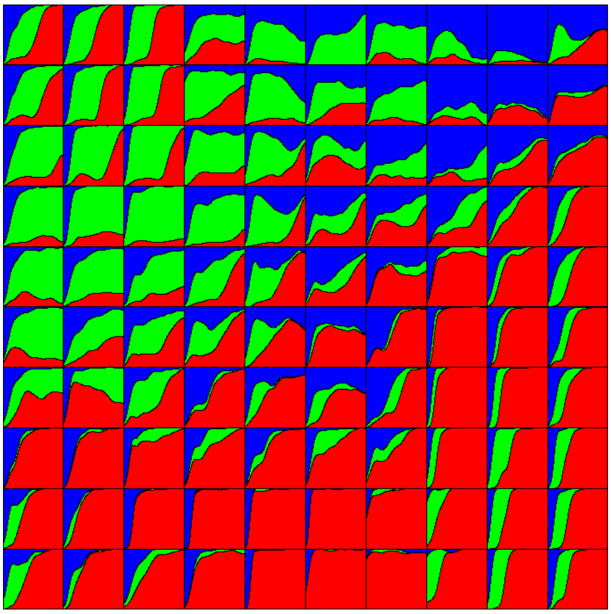
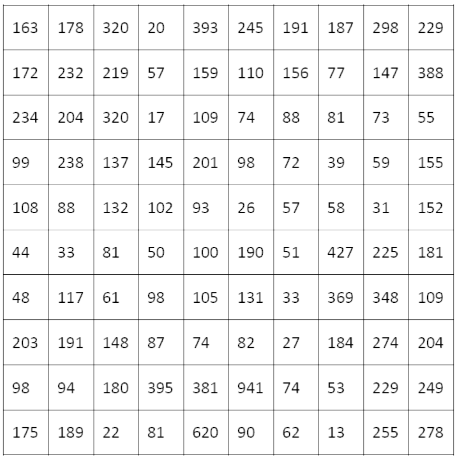


Figure 2: 10x10 - SOM



163	178	320	20	393	245	191	187	298	229
172	232	219	57	159	110	156	77	147	388
234	204	320	17	109	74	88	81	73	55
99	238	137	145	201	98	72	39	59	155
108	88	132	102	93	26	57	58	31	152
44	33	81	50	100	190	51	427	225	181
48	117	61	98	105	131	33	369	348	109
203	191	148	87	74	82	27	184	274	204
98	94	180	395	381	941	74	53	229	249
175	189	22	81	620	90	62	13	255	278

Figure 3: Number of inputs in each class

A lecture of the map summarizes the information on the career paths. Hence, we can stress out the proximities between different paths and the evolution of the career paths. Three main typologies are highlighted by the map: a stable employment position obtained relatively quickly on the *south* region of the map, a persistence in short-time jobs on the *north-west* region, and an exit from the labor market on the *north-east* region.

Thus, the last three rows and the second part of the last three columns are mostly containing permanent contracts. But these career paths are not homogeneous. One can identify some very quick transitions from school to obtaining a permanent contract (center of the last rows); these paths could be seen as the optimal school-to-work transitions. We have also some career paths starting with a short-term contract before obtaining mostly quickly a stable employment in the extreme *south-east* of the map; this trajectory corresponds to a classical path in the labor market with a first temporary job and then a stable employment. Lastly, we have other trajectories with an unemployment position after school and then a more or less rapid finding of a long-term contract in the *west* of the three last rows and in the *center* of the three last rows. These paths identify subjects with some difficulties after the obtaining their diploma but only for a short time.

The second main trajectory is concentrated in the *north-west* quarter of the map where one may find career paths with a persistence in short-term jobs. Here again there are some differences within these career paths. The 3×3 square of the *north-west* extremity is composed by a short period of unemployment, followed by a more or less long period in temporary jobs before obtaining a permanent job; hence this path leads finally to a successful integration in the labor market. The more we move to the *south* part of this quarter, the more we have persistence of temporary jobs, and at last there are some paths with a high persistence in these types of unstable contracts. At the opposite, the more we move to the *east* of the quarter, the more we find important proportions of out-of-labor-market positions with two types of career paths: an important proportion of unemployment followed by an increasing proportion of employment (stable or temporary); and a progressive exit of the labor market after a first contract.

The last main trajectory is the exclusion of the labor market that may be found in the *north-east* region of the map. There are two types of career paths: the first one is a complete and definitive exclusion of the labor market with some trajectories in constant non-employment; the second one is less problematic and is defined by a long period of exclusion of the labor market followed by an increasing ratio of employment (temporary or stable contracts). These trajectories may be found in the extreme *north-east* and in the *center* of the map. These paths illustrate some important difficulties in finding a first job after leaving school but, despite this more or less long period of unemployment there is finally a successful integration in the labor market.

4 Conclusion and future work

We have shown that clustering with SOM is useful to describe career paths and gives us some important results helping to discriminate between different typologies of employment trajectories. The use of drift Markov models allowed to take into account the impact of time on the transitions between the employment statuses.

However, there is some future work to be done. On the one hand, it would be interesting to summarize the results of the clustering into a smaller number of typologies and compute the transition probabilities within each typology. The computed transitions will give an insight on the insertion and the mobility on the labor market. This could be achieved by a hierarchical clustering on the prototypes of the map. On the other hand, the "Generation 98" survey contains variables which were not considered in our study, such as education, sex, parents' socio-professional category, geographical region. A better characterization of the career-paths should be obtained after crossing the results of the map with these variables.

References

[1] A. Abbott, A. Hrycak (1990), Measuring resemblance in sequence data: An optimal matching analysis of musician's careers, *American Journal of Sociolgy*, **vol.96(1)**, p.144-185

[2] C. Brzinsky-Fay (2007), Lost in Transition? Labour Market Entry Sequences of School Leavers in Europe, *European Sociological Review*, **vol.23(4)**, p.409-422

[3] M. Cottrell, S. Ibbou, P. Letrémy, P. Rousset (2003), Cartes auto-organisées pour l'analyse exploratoire de données et la visualisation, *Journal de la Société Française de Statistique*, **vol. 144(4)**, p.67-106

[4] D. Fougère, T. Kamionka (2005), Econometrics of individual labor market transitions, *IZA Discussion papers 1850*, Institute for the study of labor (IZA)

[5] Grelet Y., Fenelon J.-P., Houzel Y. (2000), The sequence of steps in the analysis of youth trajectories, *European Journal of Economic and Social Systems*, **vol.14(1)**

[6] J.F. Giret, P. Rousset (2007), Classifying qualitative time series with SOM : the typology of career paths in France, *Computational and ambient intelligence, IWANN 2007 Proceedings. - Edition : Berlin, Springer, 2007*, p.757-764

[7] T. Kohonen (1995) *Self Organizing Maps*, Springer-Berli

[8] S. Massoni, M. Olteanu, P. Rousset (2009), Career-path analysis using optimal matching and self-organizing maps, *Advances in Self-Organizing Maps, WSOM 2009 Proceedings. - Edition : Berlin, Springer, 2007*, p.154-162

[9] N.S. Muller, G. Ritschard, M. Studer, A. Gabadinho (2008), Extracting knowledge from life courses: Clustering and visualization, *Data Warehousing and Knowledge Discovery, 10th International Conference DaWaK 2008, Turin, Italy, September 2-5, LNCS 5182, Berlin: Springer*, p.176-185

[10] N. Vergne (2008), Drifting Markov models with polynomial drift and applications to DNA sequences, *Statistical Applications in Genetics and Molecular Biology* **vol. 7(1)**, Article 6