Career-Path Analysis Using Optimal Matching and Self-Organizing Maps

Sébastien Massoni[1], Madalina Olteanu[2], and Patrick Rousset[3]

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

Abstract. This paper is devoted to the analysis of career paths and employability. The state-of-the-art on this topic is rather poor in methodologies. Some authors propose distances well adapted to the data, but are limiting their analysis to hierarchical clustering. Other authors apply sophisticated methods, but only after paying the price of transforming the categorical data into continuous, via a factorial analysis. The latter approach has an important drawback since it makes a linear assumption on the data. We propose a new methodology, inspired from biology and adapted to career paths, combining optimal matching and self-organizing maps. A complete study on real-life data will illustrate our proposal.

1 Introduction

The question of analyzing school-to-work transitions is a challenging topic for the economists working on the labor market. In the current economic context of the world, characterized by a significant unemployment rate of young people (in France, 19.1% of the young people under 25 were unemployed during the second semester of 2008), it is interesting to study the insertion of graduates and the evolution of their career paths. The aim of this paper is to identify and analyze career-paths typologies.

Let us recall that a career path is defined as a sequence of labor-market statuses, recorded monthly. The number and the labels associated to the statuses are defined by experts. The experts take into account the fact that the more the number of labels is detailed, the more the analysis of the career paths will be accurate. We shall also remark that the labelling of the statuses is not neutral. Indeed, the choice of different criteria for identifying the statuses introduces an *a priori* separation between "good" and "bad" statuses. Generally, the number and the labelling of the statuses are quite similar in the literature, with different labels corresponding to employment and unemployment situations. The data "Generation 98" used for this study are labelled according to nine possible statuses, five employment statuses (permanent-labor contract, fixed-term contract,

J.C. Príncipe and R. Miikkulainen (Eds.): WSOM 2009, LNCS 5629, pp. 154–162, 2009.

apprenticeship contract, public temporary-labor contract, on-call contract) and four unemployment statuses (unemployed, inactive, military service, education).

In the state-of-the-art, two approaches seem to be currently used for clustering career paths. The first approach consists in transforming categorical variables into continuous variables by a factorial analysis and then apply usual clustering algorithms for Euclidean data ([9],[8]). The second approach consists in computing adapted distances for the data ([12], [4]) and then cluster using an hierarchical tree and a proximity criterion based on the distance matrix only. Both approaches have some drawbacks: in the first case, the use of factorial methods implies quite strong hypothesis concerning the linearity of the data; 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. In order to address these drawbacks, we propose to cluster career paths using a two-step methodology. The two steps of the algorithm are independent and quite general. First, we compute a dissimilarity matrix between the career paths. Second, a self-organizing map for dissimilarity matrices is trained. Besides identifying the main typologies of career paths, we are also looking for a graphical output representing the proximities and the evolutions in the different career paths.

The rest of the document is organized as follows: Section 2 is devoted to a description of the methodology and a short state-of-the-art on the subject. Section 3 contains the results on the data “Generation 98” (CEREQ, France). The conclusion and some perspectives are presented in the last section.

2 Methodology

From a statistical point of view, several problems arise when analyzing school-to-work transitions. The data sets are usually containing categorical variables, often in high dimension and have an important sample size. In order to handle these data, we made the choice of splitting the analysis into two steps. The first step consists in defining a distance or a dissimilarity measure well-suited to the data. In the second step, a clustering method is used to build and define typologies. The clustering method has to be general enough to “forget” the initial structure of the data and determine classes on the unique basis of the dissimilarity matrix computed in the previous step. This approach has the advantage of allowing a wide choice for the dissimilarity measure in the first step.

2.1 Step 1 (Optimal Matching) – Choosing a Good Distance

The first step of the analysis consists in choosing a dissimilarity measure between the career paths. Previous studies suggest the use of a multiple correspondence analysis ([9]) and a transformation of the categorical variables into continuous. This way, usual clustering algorithms based on the Euclidean distance can be trained on the factorial components. This approach has an important drawback, since it makes the assumption that data are linear. As there is no reason for this assumption to hold in our case, we prefer to use a distance which avoids this hypothesis.

Optimal matching, also known as “edit distance” or “Levenshtein distance”, was first introduced in biology by [13] and used for aligning and comparing sequences. In social sciences, the first applications are due to [1]. Let us consider two sequences of different lengths $a = (a_1, ..., a_{n_1})$ and $b = (b_1, ..., b_{n_2})$, $a_i, b_j \in S$ where S is a finite state-space. How may the two sequences be aligned and what is the distance between them? The underlying idea of optimal matching is to transform the sequence a into the sequence b using three possible operations: insertion, deletion and substitution. A cost is associated to each of the three operations. The distance between a and b is computed as the cost associated to the smallest number of operations which allow to transform a into b. The method seems simple and relatively intuitive, but the choice of the costs is a delicate operation in social sciences. This topic is subject to lively debates in the literature ([3], [14]) mostly because of the difficulties in establishing an explicit and sound theoretical frame. The interested reader may refer to [7] for a state-of-the-art.

In the data set “Generation 98”, all career paths have the same length, the status of the graduate students being observed during 94 months. Thus, we may suppose that there are no insertions or deletions and that the only cost to be computed is the substitution cost. The latter was computed using the transition matrix between the statuses as proposed in [12]: the less transitions between two statuses are observed, the more the statuses are different and the substitution cost is high. The cost w for transforming a_i into a_j is computed as a function of the observed longitudinal transitions:

$$w(a_i, a_j) = 2 - P(a_i|a_j) - P(a_j|a_i).$$

2.2 Step 2 – Self-Organizing Maps for Categorical Data

The methodology “optimal matching - clustering” has already been used ([4], [12]) during the past few years for the analysis of career or life paths. The general approach consists in computing a dissimilarity matrix using optimal matching, build an hierarchical tree and make a description of the resulting typologies. However, hierarchical clustering is limited in terms of displaying and visualizing the results. Instead, we suggest a clustering method which provides, besides clusters, a graphical representation preserving the proximity between paths.

Self-organizing maps (Kohonen algorithm, [11]) are, at the same time, a clustering algorithm and a nonlinear projection method. The input data are projected on a grid, generally rectangular or hexagonal. The grid has the important property of topology preservation: close inputs will be projected in the same class or in neighbor classes. The algorithm was initially developed for continuous data with a Euclidean distance. Since its first application, new versions, suited for particular data structures, are regularly proposed. Let us mention [10], who are using for the first time self-organizing maps and optimal matching for the analysis of biological sequences.

For career-paths data, we used a general self-organizing map algorithm, proposed by [5]. Their method does not take into account the initial structure of the data and uses only a dissimilarity matrix as input. The size and the structure of the grid must also be provided as input. At the first step of the algorithm, the prototypes are chosen at random in the input data. The rest of the algorithm consists in repeating the following steps, until the partition is stable:

- Allocating step: each input is assigned to the class of the closest prototype by minimizing a criterion of intra-class variance, extended to the neighbors. During this step, the prototypes are fixed.
- Representing step : once the new partition is determined, the new prototypes are computed by minimizing the same criterion.

The price of the generality of the method is the poverty of the space where the inputs lie: the new prototypes have to be computed and chosen between the inputs.

This algorithm is a generalization of *K-means* by introducing with a neighborhood relation defined between classes. Since the algorithm is BATCH, it is quite sensitive to the initializations.

3 Real-Life Data – "Generation 98" Survey

For illustrating the proposed methodology, we used the data in the survey "Generation 98" from CEREQ, France (http://www.cereq.fr/). 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, labelled 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 are highlighted by a first descriptive analysis of the data (Fig. 1):

- permanent-labor contract are representing 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 more 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.

Considering the important ratio of permanent contracts obtained relatively fast and its absorbing character, we decided to focus our analysis on the career paths which don't enter the "permanent contract" status in less than two years. Thus, the data set is reduced to 11777 inputs, which represent almost 3/4 of the initial data. This decision is also justified by some numerical problems such as the storage of the dissimilarity matrix and the computation time.

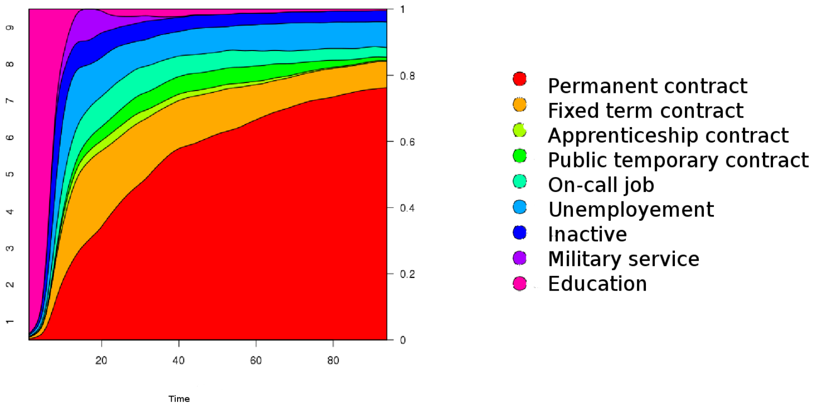

Fig. 1. Labor market structure

The inputs excluded from the analysis are grouped into a class corresponding to a "fast access to stable employment". The class contains 2032 inputs having obtained a permanent contract in less than a year and 4263 inputs having obtained a permanent contract in less than two years.

After having preprocessed the data, the analysis is conducted in three steps:

1. the transition matrix and the associated substitution-cost matrix are computed;
2. the dissimilarity matrix is computed by optimal matching with the substitution-cost matrix in step 1;
3. career paths are clustered with the self-organizing map algorithm, according to the dissimilarity matrix in step 2.

For the first two steps, we used the R-package TraMineR, available in [6]. The third step was implemented in R and is available on demand.

The cost matrix computed on the 11777 input data is the following:

$$C = \begin{pmatrix} 0 & 1.968 & 1.976 & 1.989 & 1.977 & 1.973 & 1.975 & 1.985 & 1.987 \\ 1.968 & 0 & 1.991 & 1.994 & 1.978 & 1.927 & 1.957 & 1.979 & 1.976 \\ 1.976 & 1.991 & 0 & 1.999 & 1.994 & 1.980 & 1.989 & 1.998 & 1.997 \\ 1.989 & 1.994 & 1.999 & 0 & 1.998 & 1.984 & 1.993 & 1.998 & 1.997 \\ 1.977 & 1.978 & 1.994 & 1.998 & 0 & 1.951 & 1.973 & 1.979 & 1.988 \\ 1.973 & 1.927 & 1.980 & 1.984 & 1.951 & 0 & 1.954 & 1.971 & 1.966 \\ 1.975 & 1.957 & 1.989 & 1.993 & 1.973 & 1.954 & 0 & 1.977 & 1.947 \\ 1.985 & 1.979 & 1.998 & 1.998 & 1.979 & 1.971 & 1.977 & 0 & 1.996 \\ 1.987 & 1.976 & 1.997 & 1.997 & 1.988 & 1.966 & 1.947 & 1.996 & 0 \end{pmatrix}$$

Let us remark that the values in the cost matrix are very similar. Different approaches for improving cost computation should be investigated, some perspectives are given in the conclusion.

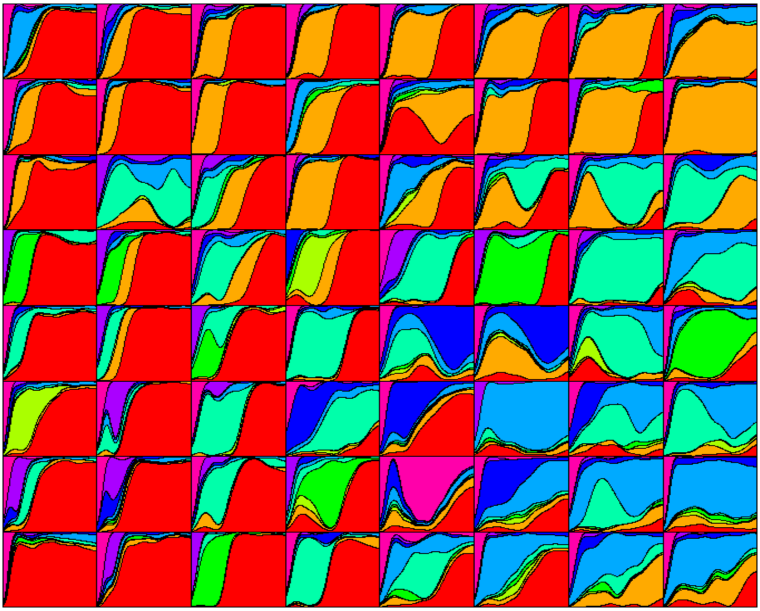

Fig. 2. SOM 8x8

When trying to compute the dissimilarity matrix, we were confronted to a numerical problem: the impossibility of storing a matrix of size 11777. Actually, this numerical problem constitutes the main drawback of the proposed methodology: the sample size must be “reasonable”. The size of the data has to be reduced before training the self-organizing map algorithm. In order to summarize the career-path data, we used a *K-modes* algorithm, which may be trained directly on the initial data set. Thus, the 11777 career-paths were summarized by 1000 representative paths and these 1000 paths were clustered with self-organizing maps.

The dissimilarity matrix between the 1000 paths was computed with the optimal matching distance and the substitution-cost matrix C. Then, we trained a self-organizing map on a rectangular grid 8×8. The resulting map is plotted in Fig.2.

A lecture of the map summarizes the information on the career paths. Thus, we can stress out the proximities between different paths and the evolution of the career paths.

The most striking opposition appears between the career paths leading to a stable-employment situation (permanent contract and/or fixed-term contract) and the career-paths more “chaotic” (unemployment, on-call contracts, apprenticeship contracts). On the map in Fig.2, the “stable” situations are mainly situated in the *west* region of the map.

Thus, the first three columns contain essentially the classes where a permanent contract was rapidly obtained. However, the *north* and the *south* regions of these columns are quite different: in the *north* region, the access to a permanent contract is achieved after a first contact with the labor-market through a fixed-term contract, while the *south* classes are only subject to transitions through a military service or education periods. At the halfway between *north* and *south* regions, we may find several apprenticeship contracts or public temporary contracts.

The stability of the career-paths noticed in the *west* region of the map is getting worse as we move to the *east*. In the *north* region, the initial fixed-term contract is getting longer until becoming a poor employment situation in the *north-east* corner. Thus, all the *east* region of the map is revealing for difficult school-to-work transitions. Let us remark the on-call contracts situations which may end by a stable contract or by unemployment. At the opposite, the career-paths starting with an apprenticeship contract are most of the time ending with a permanent contract. Finally, the *south-east* corner is characterized by exclusion career-paths: inactive and unemployment. The inactivity may appear immediately after the education period or after a first failure on the labor market.

The clustering with SOM provided interesting results by highlighting proximities or oppositions between the career paths. In order to determine a small number of typologies, we compute an hierarchical classification tree on the 64 prototypes of the map. Ward criterion was used for aggregating classes. The clustering is represented in Fig.3. The final configuration with nine classes allows to describe 8 career-path typologies:

1. relatively fast access to a stable situation (*class 1*)
2. transition through a fixed-term contract before obtaining a permanent contract(*class 2*)
3. fixed-term contracts (*class 3*)

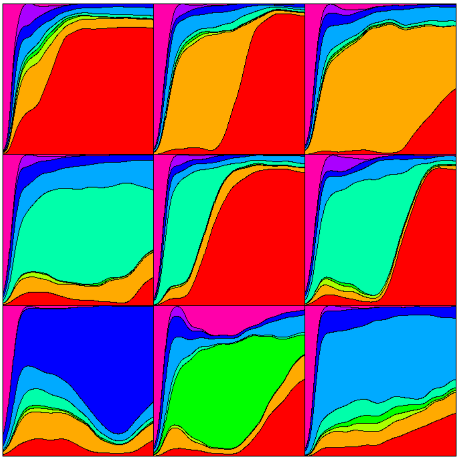

Fig. 3. Final career-path typologies

4. on-call contracts with periods of unemploymentnt (*class 4*)
5. shorter or longer on-call contracts and finally a permanent contract(*classes 5 and 6*)
6. inactivity after loosing a first job or immediately after school (*class 7*)
7. apprenticeship contract ending with a fixed-term or a permanent-term contract(*class 8*)
8. long unemploymentnt period with a gradual return to employment (*class 9*)

We may also add the first class which was excluded from the analysis:

9. fast access to stabl employment after leaving school.

The relative importance of every class is given in Table 1.

Table 1. Importance of final classes

Class	Size	Weight in the sample	Weight in the whole sample
1.	5475	46.5%	34.1%
2.	968	8.2%	6.0%
3.	1082	9.2%	6.8%
4.	514	4.4%	3.2%
5.	1122	9.5%	7.0%
6.	328	2.8%	2.0%
7.	1002	8.5	6.3%
8.	1286	10.9	8.0%
9.	4263	NA	26.6%

4 Conclusion and Perspectives

Several typologies for career paths were highlighted by our analysis. The self-organizing map allowed a detailed characterization of the proximities, oppositions and transitions between the different career paths. The proposed methodology is thus well suited for this kind of data.

However, several aspects should be improved in this approach. Concerning the labelling of the statuses, we remarked that the “military service” is not representative in a employment-unemployment discrimination. The solution is either to change the label of this status, either to delete this period. In the latter case, we would no longer have equal-sized sequences and the question of computing an insertion/deletion cost would arise.

The second remark concerns the computation of the substitution-cost matrix. In this paper, it was computed using the observed transitions, but without taking into account that these transitions change in time. Indeed, in a more realistic frame we should consider the non-homogeneity of the transitions and probably use non-homogeneous Markov chains in order to estimate the costs.

References

1. Abbott, A., Forrest, J.: Optimal matching methods for historical sequences. Journal of Interdisciplinary History 16, 471–494 (1986)
2. Abbott, A., Hrycak, A.: Measuring resemblance in sequence data: An optimal matching analaysis of musicians carrers. American Journal of Sociolgy 96(1), 144–185 (1990)
3. Abbott, A., Tsay, A.: Sequence analysis and optimal matching methods in sociology. Review and prospect. Sociological Methods and Research 29(1), 333 (2000)
4. Brzinsky-Fay, C.: Lost in Transition? Labour Market Entry Sequences of School Leavers in Europe. European Sociological Review 23(4), 409–422 (2007)
5. Conan-Guez, B., Rossi, F., El Golli, A.: Fast algorithm and implementation of dissimilarity Self-Organizing Maps. Neural Networks 19(6-7), 855–863 (2006)
6. Gabadinho, A., Ritschard, G., Studer, M., Müller, N.S.: Mining Sequence Data in R with TraMineR: a user's guide (2008), `http://mephisto.unige.ch/traminer`
7. Gauthier, J.A., Widmer, E.D., Bucher, P., Notredame, C.: How much does it cost? Optimization of costs in sequence analysis in social science data. Sociological Methods and Research (in press) (2008)
8. Giret, J.F., Rousset, P.: Classifying qualitative time series with SOM: the typology of career paths in France. In: Computational and ambient intelligence, IWANN 2007 Proceedings, pp. 757–764. Springer, Berlin (2007)
9. Grelet, Y., Fenelon, J.-P., Houzel, Y.: The sequence of steps in the analysis of youth trajectories. European Journal of Economic and Social Systems 14(1) (2000)
10. Kohonen, T., Somervuo, P.: Self-organizing maps for symbol strings. Neurocomputing 21, 19–30 (1998)
11. Kohonen, T.: Self Organizing Maps. Springer, Berlin (1995)
12. Müller, N.S., Ritschard, G., Studer, M., Gabadinho, A.: Extracting knowledge from life courses: Clustering and visualization. In: Song, I.-Y., Eder, J., Nguyen, T.M. (eds.) DaWaK 2008. LNCS, vol. 5182, pp. 176–185. Springer, Heidelberg (2008)
13. Needleman, S., Wunsch, C.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48(3), 443–453 (1970)
14. Wu, L.: Some comments on Sequence analysis and optimal matching methods in sociology, review and prospect. Sociological methods and research 29(1), 41–64 (2000)