Career-path analysis using Drifting Markov Models and Self-Organizing Maps

Sébastien MASSONI[1],
Madalina OLTEANU[2], and Patrick ROUSSET[3]

[1] CES - Université Paris 1 & Paris School of Economics
[2] SAMM - Université Paris 1
[3] CEREQ

Workshop on "Trajectories" - October 14, 2011

Outline

Employability and career-path analysis

- Unemployment rate among young under 25 in France
  - 23% in the beginning of 2011
  - between 16 and 25% since 1980
- Analysis of school-to-work transitions
- Identify the main career-path typologies
  (stable, unstable, insecurity, out of the market,...)

Employability

Process allowing graduates to reach a stable job or position
(Vernières, 1997)

Career path

Vector of monthly situations of an individual on the labor market

Clustering career-paths

- Categorical data in high dimension

Current approaches:

- Multiple correspondence analysis, followed by any clustering method for Euclidean data (Grelet et al., 2000)
  - + Categorical variables are transformed into continuous variables
  - - The Euclidean distance may not be suited for the data by assuming strong hypothesis of linearity
- Specific dissimilarities for employability profiles, followed by hierarchical clustering (Muller et al., 2008)
  - + Large choice for the dissimilarity measure
  - - Hierarchical clustering does not provide a tool for visualizing results and is not suited for large datasets

Clustering career-paths

- Categorical time series

Current approaches:

- Optimal matching to compute the distance between two paths (Massoni et al., 2009)
  - \+ Take into account dependence between variables
  - \- Transition, insertion, and deletion costs have to be defined a priori
- First-order Markov chains to model the career-paths (Fougère and Kamionka, 2008)
  - \+ Transform paths into quantitative data
  - \- Homogeneous Markov chains assume same probabilities of transition

Proposed methodology

Two independent steps for clustering career-paths (categorical time-series of different lengths):

- Estimate transition probabilities for every path with a non-homogeneous Markov chain
  $\Rightarrow$ Drifting Markov Models (DMM)
- Run a clustering algorithm designed for quantitative data with missing values on theses estimates
  $\Rightarrow$ Self-Organizing Maps (SOM)

Goals:

- Identify career-path typologies
- Visualize career-paths, highlight proximities and oppositions between profiles

Markov chains: basics

Random process that define probabilities of transitions from one state to another without memory (the next state depends only on the current state)

$$\mathbb{P}(X_{n+1} = x | X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = \mathbb{P}(X_{n+1} = x | X_n = x_n)$$

- Example: weather forecast with two states *sun* and *rain*
  - A sunny day is 90% likely to be followed by another sunny day, and a rainy day is 50% likely to be followed by another rainy day
  - The transition matrix P is $\begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}$
  - The weather on day 0 is known to be sunny: $\mathbf{x}^{(0)} = [1 \ \ 0]$
  - The weather on day 1 can be predicted by:
    $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} P = [1 \ \ 0] \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix} = [0.9 \ \ 0.1]$
  - The weather on day 2 can be predicted in the same way:
    $\mathbf{x}^{(2)} = \mathbf{x}^{(1)} P = \mathbf{x}^{(0)} P^2 = [1 \ \ 0] \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}^2 = [0.86 \ \ 0.14]$
  - General rules for day n are: $\mathbf{x}^{(n)} = \mathbf{x}^{(n-1)} P = \mathbf{x}^{(0)} P^n$

Non-Homogeneous Markov chains

- A non-homogeneous Markov chain (variation of the transition matrix) will be associated to each career-path
- The estimates of the transition probabilities will characterize the corresponding career-path and then can be used as clustering variables in the SOM

Advantages:

- Dependence between the variables in the path
- Impact of time on the transitions between statuses

Methods:

- Hidden Markov Models
  $\Rightarrow$ Different transition matrices are fitted on different homogeneous parts of the trajectory
- Drifting Markov Models (Vergne, 2008)
  $\Rightarrow$ An initial and a final transition matrix linked together by a linear or a polynomial deterministic function

DMM with linear drift

Consider $X = (X_t)_{t \in \{0,\dots,n\}}$ a process defined by state space $\mathcal{A}$ and a transition matrix

$$\pi_{\frac{t}{n}}(u, v) = \mathbb{P}(X_t = v | X_{t-1} = u) \ , \ \forall \ u, v \in \mathcal{A}$$

- The parameters defining the model are π_0 and π_1.
- Linear drift assumes $\pi_{\frac{t}{n}}$ as a linear combination of the initial and the final transition matrices, π_0 and π_1:

$$\pi_{\frac{t}{n}} = \left(1 - \frac{t}{n}\right)\pi_0 + \frac{t}{n}\pi_1$$

DMM with linear drift - estimation procedure

The estimates of π_0 and π_1 are computed by minimizing the sum of prediction errors:

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

Explicit form for the estimates 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

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

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

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

Self-Organizing Maps

Data clustering (vector quantization) and nonlinear projection

- The inputs' data is projected onto a two-dimensional grid
- The projection respects the topology of the data
  $\Rightarrow$ two vectors closed in initial space will be projected in same or neighbor classes

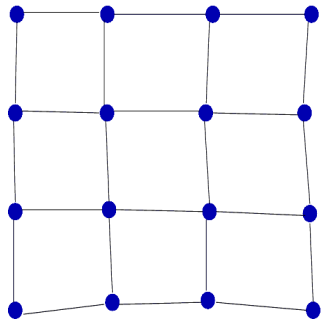
Main advantage of clustering via SOM is to provide more information by given:

- An insight on the proximities of the clusters
- A two-dimensional representation of the data

Idea of SOM

Based on inputs $\Omega = \{x_1, x_2, ..., x_n\}$, a priori choices of:

- Number of classes C
- Map structure: (C, Γ)

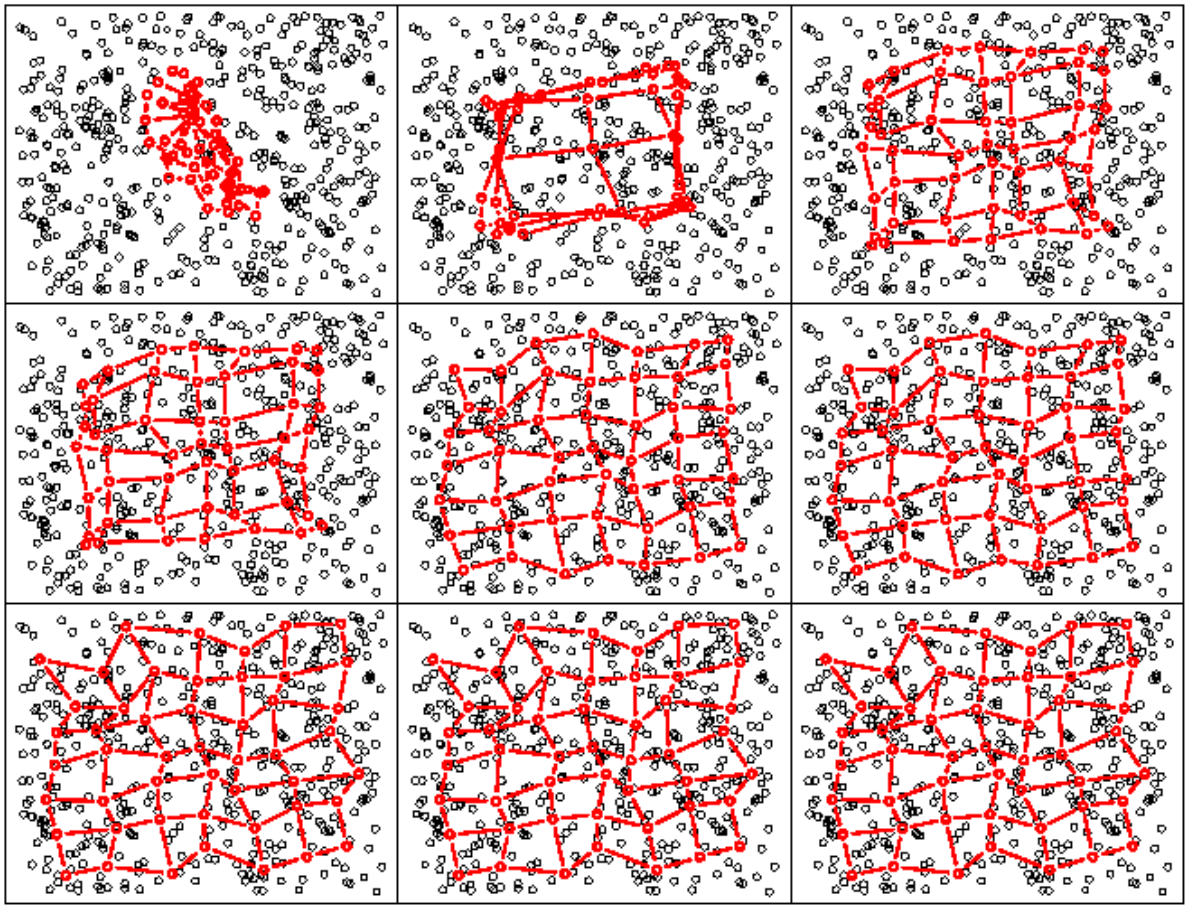


- Neighborhood function between classes: $K : \mathbb{R}_+ \to \mathbb{R}_+$, $K(0) = 1$, $\lim_{x \to \infty} K(x) = 0$
  - usually, $K(x) = \exp\left(-x^2\right)$
- Minimal distance between two neurons c and c': $\delta\left(c, c'\right)$

$\Rightarrow$ The idea is to associate to each neuron c a vector code p_c and a subset $C_c \in \Omega$ and then to minimize the intra-classes variance.

Illustration

1000 points of a Uniform distribution on $[0, 1]^2$ on a 7×7-grid



Kohonen algorithm

- To each neuron c of the map C a code-vector p_c and a subset C_c of the data will be associated
- Cost-function:

$$E\left(\left(C_c\right)_{c\in C},\left(p_c\right)_{c\in C}\right)=\sum_{x_i\in\Omega}\sum_{c\in C}K^T\left(\delta\left(f\left(x_i\right),c\right)\right)d^2\left(p_c,x_i\right)$$

2 steps:

- Allocation: closest vector-code for each observation

$$f(x_i)=\arg\min_{r\in C}\gamma^T\left(x_i,r\right)$$
$$\gamma^T\left(x_i,r\right)=\sum_{c\in C}K^T\left(\delta\left(r,c\right)\right)d^2\left(x,c\right)$$

- Representation: compute vector-code for each class

$$p_c=\arg\min_{p\subset\Omega,|p|=q}E_c^T\left(p\right)$$
$$E_c^T\left(p\right)=\sum_{x_i\in\Omega}K^T\left(\delta\left(f(x_i),c\right)\right)d^2\left(p_c,x_i\right)$$

SOM with missing values

Kohonen algorithm perfectly deals with data with missing values without estimating them beforehand

- If one status is missing from the career-path $\Rightarrow$ transition probabilities associated will not be estimated
- SOM algorithm has to handle the case where several values in an input vector are missing

$\Rightarrow$ Approach used: the winning prototype for each input is computed only on the available variables (Cottrell et al., 2003)

"Generation 98" data set

- 16040 graduates in 1998
- Monthly statuses recorded during 94 months
- Nine possible statuses :
  - Permanent contract,
  - Fixed-term contract, apprenticeship contract, public temporary contract, on-call job,
  - Unemployment, inactive, education,
  - Military service

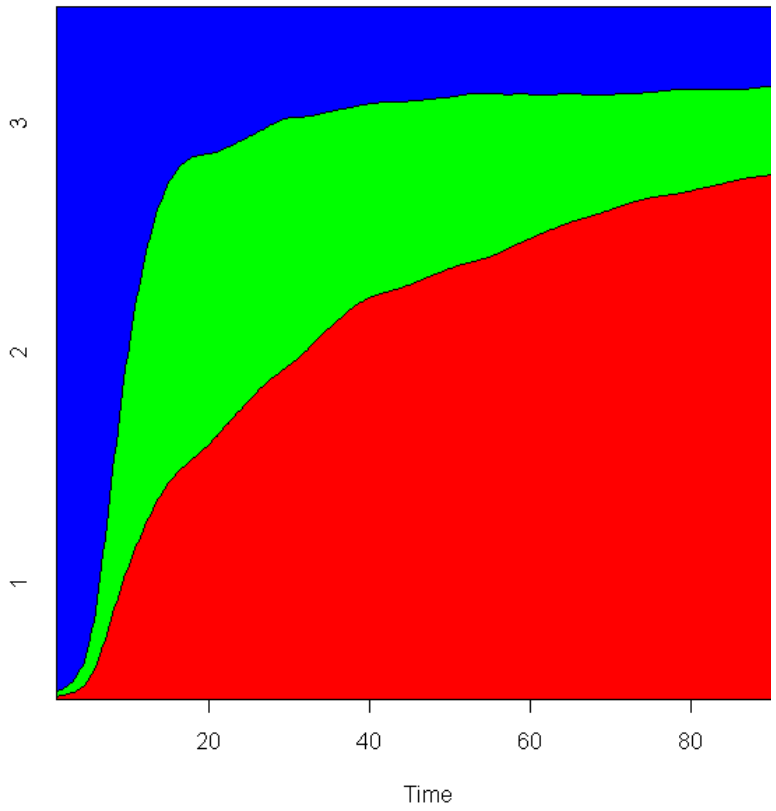
Descriptive statistics

- Permanent-contract ratio: 20% (one year); 50% (three years); 75% (seven years)
- Fixed-term contract ratio: 20% (one year); 15% (three years); 8% (seven years)
- Unemployment ratio: 30% (one year); 11% starting with the fourth year

Two transformations were made in the original data set:

- The "military service" status was suppressed (no information on employment status)
- Number of categories was reduced to three: "stable employment", "short-time employment", and "out-of-the-labor-market"

Labor market structure



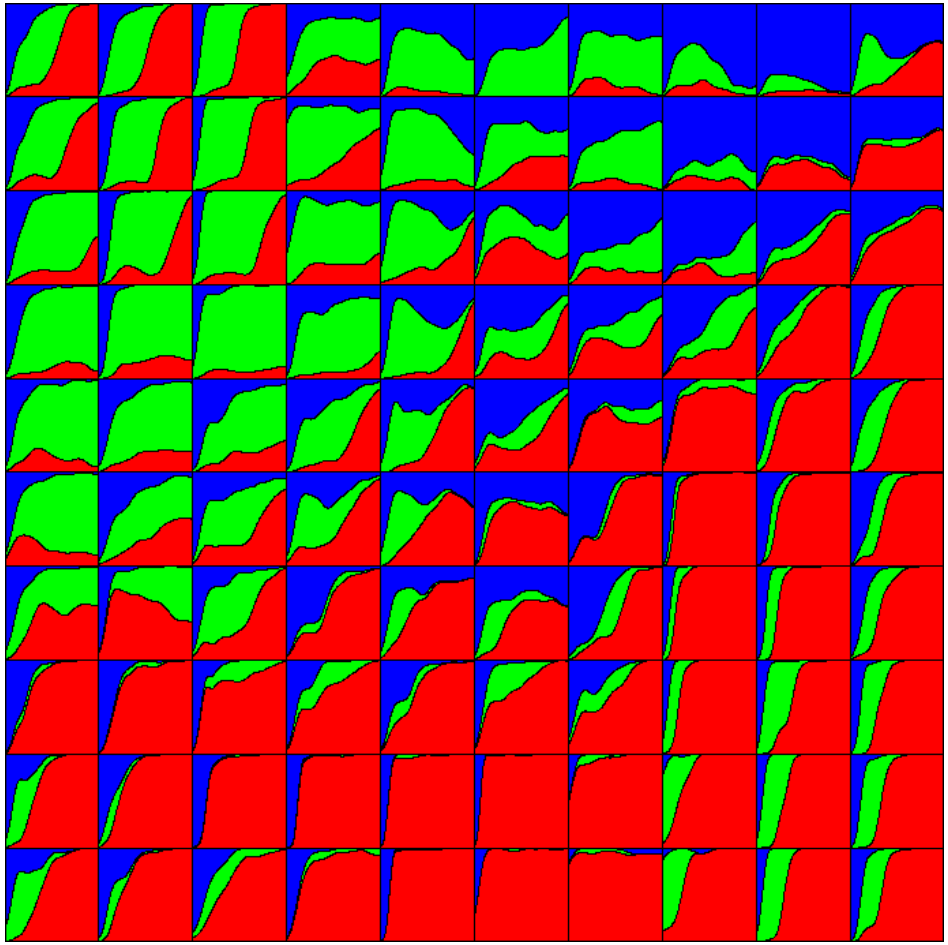
Clustering algorithm

Two steps:

1. Estimate the DMM corresponding to each career-path
   - To each input vector correspond two estimated transition matrices with elements stored as vector of size 18
   - The initial 16040x94 categorical data set was transformated into a 16040x18 unnormalized quantitative data set

2. Cluster career-paths using SOM
   - Modified to handle missing values
   - Using Euclidean distance
   - Projection structure on a 10 x 10 rectangular grid
   - SAS programs developed by Letremy

Results: 10 x 10 SOM



Results: class size

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

Lecture of the map

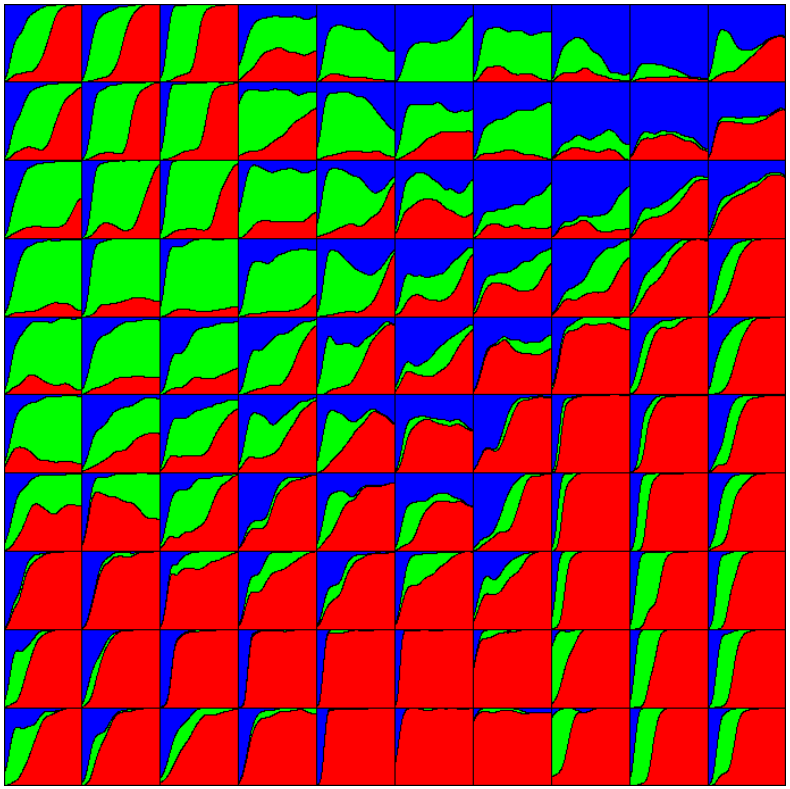
Identify proximities between different paths and evolution of career-paths

$\Rightarrow$ Three main typologies:

- *South*: a stable employment position obtained quickly
- *North-West*: a persistence in short-time jobs
- *North-East*: an exit from the labor market

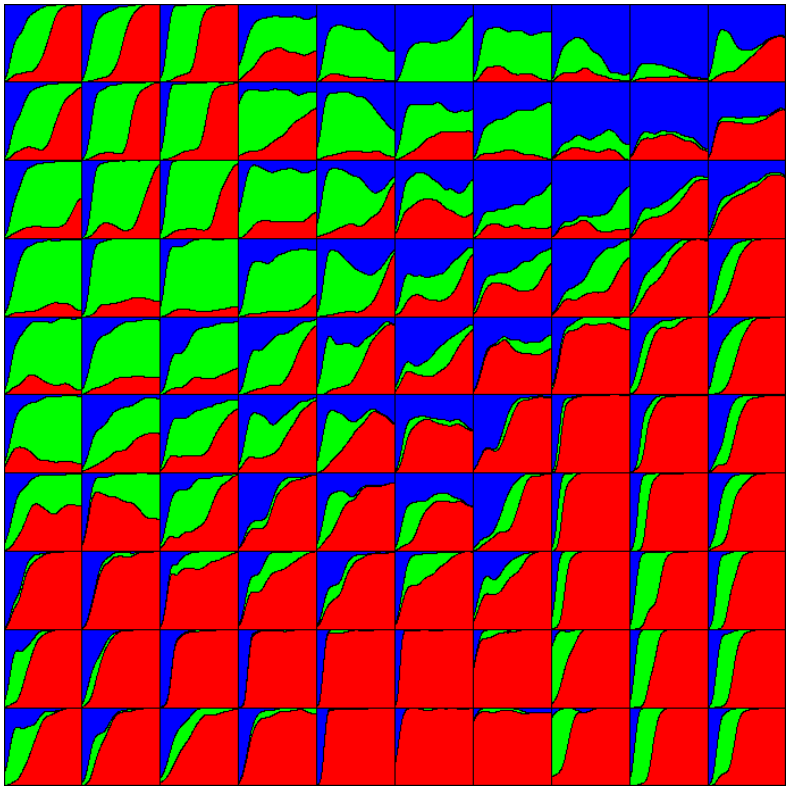
But in each of theses typologies we have some heterogeneity in the career-paths

Stable employment (*south*)



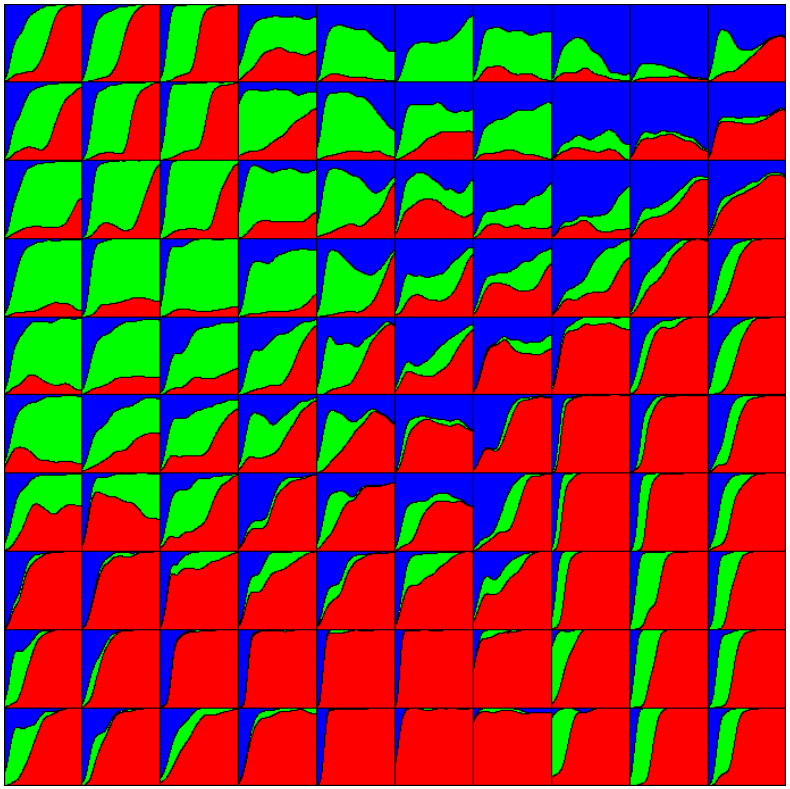
- School ⇒ Long-term contract (*center*)
- Short-term contract ⇒ Long-term contract (*east*)
- Unemployment ⇒ Long-term contract (*west*)

Short-time job (*north-west*)



- Unemployment $\Rightarrow$ Short-term contract $\Rightarrow$ Long-term contract (*west*)
- Unemployment $\Rightarrow$ Short-term contract (*south*)
- Long unemployment $\Rightarrow$ Employment or exit (*south*)

Exit from the labor market (*north-east*)



- Unemployment $\Rightarrow$ Exclusion (*east*)
- Long unemployment $\Rightarrow$ Employment (*center*)

Conclusion

- SOM provides a visualization tool for the proximities and oppositions between career-paths
- DMM allows to take into account the impact of time on the transitions between the employment statuses

Future works

- Reduce the number of categories: *with* or *without* employment
  $\Rightarrow$ Avoid the problem of missing data and increase the number of observations per parameters

- Compute the χ^2 distance instead of the euclidian one in the SOM algorithm
  $\Rightarrow$ Weighting the differents status

- Study the transition matrices within each typology
  $\Rightarrow$ Hierarchical clustering on the prototypes of the map to obtain macro-classes

- Better characterization of the career-paths
  $\Rightarrow$ Crossing the results of the map with socio-economics variables (education, sex, parents' socio-professional category, ...)