Which dissimilarity is to be used when extracting typologies in sequence analysis? A comparative study

Sébastien Massoni, Madalina Olteanu, Nathalie N. Villa-Vialaneix

Which dissimilarity is to be used when extracting typologies in sequence analysis? A comparative study

Sébastien Massoni[1], Madalina Olteanu[2], and Nathalie Villa-Vialaneix[2,3]

[1] Centre d'Economie de la Sorbonne, UMR CNRS 8174, Université Paris 1
sebastien.massoni@gmail.com
[2] SAMM, EA 4543, Université Paris 1, Paris, France
madalina.olteanu@univ-paris1.fr
[3] Unité MIAT, INRA de Toulouse, Auzeville, France
nathalie.villa@toulouse.inra.fr

Abstract. Originally developed in bioinformatics, sequence analysis is being increasingly used in social sciences for the study of life-course processes. The methodology generally employed consists in computing dissimilarities between the trajectories and, if typologies are sought, in clustering the trajectories according to their similarities or dissemblances. The choice of an appropriate dissimilarity measure is a major issue when dealing with sequence analysis for life sequences. Several dissimilarities are available in the literature, but neither of them succeeds to become indisputable. In this paper, instead of deciding upon one dissimilarity measure, we propose to use an optimal convex combination of different dissimilarities. The optimality is automatically determined by the clustering procedure and is defined with respect to the within-class variance.

1 Introduction

Originally developed in bioinformatics, sequence analysis is being increasingly used in social sciences for the study of life-course processes. The methodology generally employed consists in computing dissimilarities between the trajectories and, if typologies are sought, in clustering the trajectories according to their similarities or dissemblances. However, measuring dissimilarities or similarities for categorical sequences has always been a challenge in practice. This challenge becomes even harder in social sciences where these measures need some theoretical foundations. Choosing the appropriate dissimilarity or dissimilarity for life-sequence analysis is a key issue which relates to the resulting typologies. The literature on this topic is very rich and still very debated. Each method has its own advantages and drawbacks [1, 2].
In this paper, we introduce a different approach. Instead of deciding upon one specific dissimilarity, we propose to use several ones, optimally combined. We consider three main categories of dissimilarities : χ^2-metric [3], optimal matching [2] and non-alignement techniques [1]. Since our final goal is to extract typologies for life sequences, we are looking for

the best convex combination of the different dissimilarities which provides the best clusters in terms of homogeneity. The algorithm used for clustering is a self-organizing map (SOM). We use a modified version of the online relational SOM introduced in [4]. In the algorithm proposed here, an additional step is added to each iteration. During this step, the coefficients of the convex combination of dissimilarities are updated according to a gradient-descent principle which aims at minimizing the extended within-class variance.
The rest of the manuscript is organized as follows : Section 2 reviews the different dissimilarities usually used to handle categorical time series. Section 3 describes the online relational SOM for multiple dissimilarities. Section 4 presents a detailed application for sequences related to school-to-work transitions.

2 Dissimilarities for Life Sequences

Three main categories of dissimilarities were addressed in our study. Each of them is briefly described below.
χ^2-distance. Historically, factor analysis was used first to extract typologies from life sequences, [3]. The sequences, which are categorical data, were transformed by running a multiple correspondence analysis (MCA) on the complete disjunctive table. Then, clustering methods adapted to continuous data were applied and the main typologies were extracted. Performing MCA and then computing the Euclidean distance on the resulting vectors is equivalent to computing the χ^2-distance on the rows of the complete disjunctive table. The χ^2-distance is weighting each variable by the inverse of the associated frequency. Hence, the less frequent situations have a larger weight in the distance and the rare events become more important. Also, the χ^2-distance emphasizes the contemporary identical situations, whether these identical moments are contiguous or not. However, the transitions between statuses are not taken into account and input vectors are close only if they share contemporary statuses throughout time.
Optimal-matching dissimilarities. Optimal matching, also known as "edit distance" or "Levenshtein distance", was first introduced in biology by [5] and used for aligning and comparing sequences. In social sciences, the first applications are due to [6]. The underlying idea of optimal matching is to transform the sequence i into the sequence i' using three possible operations: insertion, deletion and substitution. A cost is associated to each of the three operations. The dissimilarity between i and i' is computed as the cost associated to the smallest number of operations which allows to transform i into i'. 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 [7, 8] mostly because of the difficulties to establish an explicit and sound theoretical frame. Among optimal-matching dissimilarities, we selected three dissimilarities: the OM with substitution costs computed from the transition matrix between statuses as proposed in [9], the Hamming dissimilarity (HAM, no insertion or deletion costs and a

substitution cost equal to 1) and the Dynamic Hamming dissimilarity (DHD as described in [10]). Obviously, other choices are equally possible and the costs may be adapted, depending whether the user wants to highlight the contemporaneity of situations or the existence of common, possibly not contemporary, sub-sequences.

Non-alignment techniques. Since the definition of costs represents an important drawback for optimal-matching dissimilarities, several alternatives were proposed in the literature. Here, we considered three different dissimilarities introduced by C. Elzinga [1, 11]: the longest common prefix (LCP), the longest common suffix or reversed LCP (RLCP) and the longest common subsequence (LCS). Dissimilarities based on common subsequences are adapted to handle transitions between statuses while they take into account the order in the sequence. They are also able to handle sequences of different lengths.

3 Relational SOM

Extracting typologies from life sequences requires clustering algorithms based on dissimilarity matrices. Generally, hierarchical clustering or K-means are used in the literature, [2]. In this paper, we focus on a different approach, based on a Self-Organizing Map (SOM) algorithm [12]. The interest of using a SOM algorithm adapted to dissimilarity matrices was shown in [13]. Self-organizing maps possess the nice property of projecting the input vectors in a two-dimensional space, while clustering them. In [13], the authors used dissimilarity SOM (DSOM) introduced by [14]. OM with substitution cost defined from the transition matrix was used to measure the dissimilarity between sequences. While DSOM improves clustering by additionally providing a mapping of the typologies, it still has a major drawback: prototypes have to be chosen among the input vectors. Thus, the clustering doesn't allow for empty clusters, which may be quite restrictive in some cases. Moreover, this property of DSOM makes it very sensitive to the initialization. The computation time is also very important, since the research of the prototype is done exhaustively among all input vectors and the algorithm is of batch type.

Online relational SOM. Inspired by the online kernel version of SOM [15], [4] recently proposed an online version of SOM for dissimilarity matrices, called online relational SOM. Online relational SOM is based on the assumption that prototypes may be written as convex combinations of the input vectors, as previously proposed in [16]. This assumption gives more flexibility to the algorithm, which now allows for empty clusters. Moreover, since the algorithm is online, the dependency on the initialization lessens and the computation time also decreases.

In the online relational SOM, n input data, $x_1, \ldots, x_n$, taking values in an arbitrary input space $\mathcal{G}$, are described by a dissimilarity matrix $\mathbf{\Delta} = (\delta_{ij})_{i,j=1,\ldots,n}$ such that $\mathbf{\Delta}$ is non negative ($\delta_{ij} \geq 0$), symmetric ($\delta_{ij} = \delta_{ji}$) and null on the diagonal ($\delta_{ii} = 0$). The algorithm maps the data into a low dimensional grid composed of U units which are linked together by a neighborhood relationship $H(u, u')$. A prototype p_u is associated with each unit $u \in \{1, \ldots, U\}$ in the grid. To allow computation

of dissimilarities between the prototypes $(p_u)_u$ and the data $(x_i)_i$, the prototypes are symbolically represented by a convex combination of the original data $p_u \sim \sum_i \beta_{ui} x_i$ with $\beta_{ui} \in [0,1]$ and $\sum_i \beta_{ui} = 1$.

Online multiple relational SOM. As explained in the introduction, the choice of a dissimilarity measure in social sciences is a complex issue. When the purpose is to extract typologies, the results of the clustering algorithms are highly dependent on the criterion used for measuring the dissemblance between two sequences of events. A different approach is to bypass the choice of the metric: instead of having to choose one dissimilarity measure among the existing ones, use a combination of them. However, this alternative solution requires an adapted clustering algorithm.

Similarly to the multiple kernel SOM introduced in [17], we propose the multiple relational SOM (MR-SOM). Here, D dissimilarity matrices measured on the input data, $\mathbf{\Delta}^1, \ldots, \mathbf{\Delta}^D$, are supposed to be available. These matrices are combined into a single one, defined as a convex combination: $\mathbf{\Delta}^\alpha = \sum_d \alpha_d \mathbf{\Delta}^d$ where $\alpha_d \geq 0$ and $\sum_{d=1}^D \alpha_d = 1$.

If the (α_d) are given, relational SOM based on the dissimilarity $\mathbf{\Delta}^\alpha$ aims at minimizing over $(\beta_{ui})_{ui}$ and $(\alpha_d)_d$ the following energy function :

$$\mathcal{E}((\beta_{ui})_{ui}, (\alpha_d)_d) = \sum_{i=1}^{n} \sum_{u=1}^{U} H\left(f(x_i), u\right) \delta^\alpha\left(x_i, p_u(\beta_u)\right) ,$$

where $f(x_i)$ is the neuron where x_i is classified[4], $\delta^\alpha\left(x_i, p_u(\beta_u)\right)$ is defined by $\delta^\alpha\left(x_i, p_u(\beta_u)\right) \equiv \mathbf{\Delta}_i^\alpha \beta_u - \frac{1}{2}\beta_u^T \mathbf{\Delta}^\alpha \beta_u$ and $\mathbf{\Delta}_i^\alpha$ is the *i-th* row of the matrix $\mathbf{\Delta}^\alpha$.

When there is no a-priori on the $(\alpha_d)_d$, we propose to include the optimization of the convex combination into an online algorithm that trains the map. Following an idea similar to that of [18], the SOM is trained by performing, alternatively, the standard steps of the SOM algorithm (i.e., affectation and representation steps) and a gradient descent step for the $(\alpha_i)_i$. To perform the stochastic gradient descent step on the (α_d), the computation of the derivative of $\mathcal{E}|_{x_i} = \sum_{u=1}^{M} H\left(f(x_i), u\right) \delta^\alpha\left(x_i, p_u(\beta_u)\right)$ (the contribution of the randomly chosen observation $(x_i)_i$ to the energy) with respect to α is needed. But, $\mathcal{D}_{id} = \frac{\partial \mathcal{E}|_{x_i}}{\partial \alpha_d} = \sum_{u=1}^{M} H\left(f(x_i), u\right) \left(\mathbf{\Delta}_i^d \beta_u - \frac{1}{2}\beta_u^T \mathbf{\Delta}^d \beta_u\right)$, which leads to the algorithm described in Algorithm 1.

4 Application for the analysis of Life Sequences

Data. For illustrating the proposed methodology and its relevance for categorical time series analysis, we used the data in the survey "Generation 98" from CEREQ, France (http://www.cereq.fr/). According to the French National Institute of Statistics, 22,7% of young people under 25 were unemployed at the end of the first semester 2012.[5] Hence,

[4] Usually, it is simply the neuron whose prototype is the closest to x_i: see Algorithm 1.

[5] All computations were performed with the free statistical software environment **R** (`http://cran.r-project.org/`, [19]). The dissimilarity matrices (except for the χ^2-

Algorithm 1 Online multiple dissimilarity SOM

1: $\forall\, u$ and i initialize β_{ui}^0 randomly in $\mathbf{R}$ and $\forall\, d$, set α_d.
2: **for** t=1,. . . ,T **do**
3: Randomly choose an input x_i
4: *Assignment step*: find the unit of the closest prototype

$$f^t(x_i) \leftarrow \arg \min_{u=1,\ldots,M} \delta^{\alpha,t}\left(x_i, p_u(\beta_u)\right)$$

5: *Representation step*: update all the prototypes: $\forall\, u$,

$$\beta_{ul}^t \leftarrow \beta_{ul}^{t-1} + \mu(t) H(f^t(x_i), u)\left(\delta_{il} - \beta_{ul}^{t-1}\right)$$

6: *Gradient descent step*: update the dissimilarity: $\forall\, d = 1, \ldots, D$,

$$\alpha_d^t \leftarrow \alpha_d^{t-1} + \nu(t)\mathcal{D}_d^t \qquad \text{and} \qquad \delta^{\alpha,t} \leftarrow \sum_d \alpha_d^t \delta^d.$$

7: **end for**

the question of how is achieved the transition from school to employment or unemployment is crucial in the current economic context. The dataset contains information on 16 040 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 are highlighted by a first descriptive analysis of the data as shown in Figure 1:

- 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 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.

In this dataset, all career paths have the same length, the status of the graduate students being observed during 94 months. Hence, we suppose that there are no insertions or deletions and that only the substitution costs have to be defined for OM metrics. This is equivalent to supposing low substitution costs with respect to the insertion-deletion costs. This choice may be considered restrictive, since in this case the OM metrics will only highlight the contemporaneity of situations. However, Elzinga metrics such as the LCS used in the manuscript are built starting from common, although not contemporary, subsequences and are very sim-

distance) and the graphical illustrations were carried out using the `TraMineR` package [20]. The online multiple dissimilarity SOM was implemented by the authors.

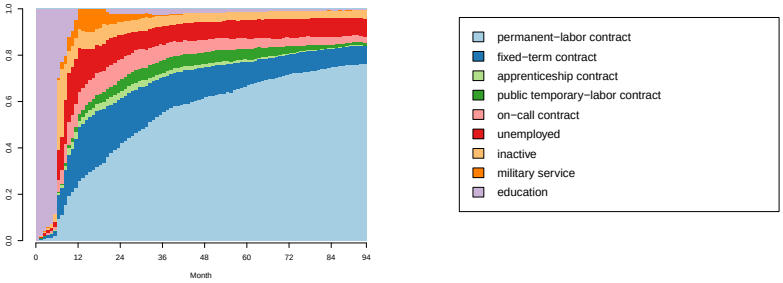

Fig. 1. Labor market structure

ilar to OM dissimilarities with insertion-deletion costs lower than the substitution costs.
Seven different dissimilarities were considered: the χ^2-distance, the Hamming dissimilarity (HAM), OM with substitution-cost matrix computed from the transition matrix as shown in Section 2, the dynamic Hamming dissimilarity (DHD) as defined in [10], the longest common prefix (LCP), the longest common suffix or reversed LCP (RLCP), the longest common substring (LCS).

5 Preliminary study

Since the original data contain more than 16 000 input sequences and since the relational SOM algorithms are based on dissimilarity matrices, the computation time becomes rapidly very important. Training the map on the entire data set requires several hours or days of computing time. Hence, in order to identify the role of the different dissimilarities in extracting typologies, we considered several samples drawn at random from the data. For each of the experiments below, 50 samples containing 1 000 input sequences each were considered. For each sample, the seven dissimilarity matrices listed above were computed and normalized according to the max norm. In order to assess the quality of the maps, three indexes were computed : the quantization error and the dispersion between prototypes for quantifying the quality of the clustering and the topographic error for quantifying the quality of the mapping, [21]. These quality criterai all depend on the dissimilarities used to train the map but the results are made comparable by using normalized dissimilarities.
Optimal-matching metrics. The first experiment was concerned with the three optimal-matching metrics. The results are listed in Table 1. According to the mean values of the α's, the three dissimilarities contributed to extracting typologies. The Hamming and the dynamical Hamming dissimilarities have similar weights, while the OM with cost-matrix defined from the transition matrix has the largest weight. The mean quantization error computed on the maps trained with the three dissimilarities

optimally combined is larger than the quantization error computed on the map trained with the OM metric only. On the other hand, the topographic error is improved in the mixed case. In this case, the joint use of the three dissimilarities provides a trade-off between the quality of the clustering and the quality of the mapping. The results in Table 1 confirm the difficulty to define adequate costs in OM and the fact that the metric has to be chosen according to the aim of the study : building typologies (clustering) or visualizing data (mapping).

a) Optimally-tuned α

Metric	OM	HAM	DHD
α-Mean	0.43111	0.28459	0.28429
α-Std	0.02912	0.01464	0.01523

b) Quality criteria for the SOM-clustering

Metric	OM	HAM	DHD	Optimally-tuned α
Quantization error	92.93672	121.67305	121.05520	114.84431
Topographic error	0.07390	0.08806	0.08124	0.05268
Prototype dispersion	2096.95282	2255.36631	2180.44264	2158.54172

Table 1. Preliminary results for three OM metrics

Elzinga metrics. When MR-SOM clustering is performed using the three Elzinga metrics only, the results in Table 2 are clearly in favor of the LCS. This result is less intuitive. For example, the LCP metric has been widely used in social sciences and more particularly for studying school-to-work transitions. Indeed, it is obvious that all sequences start with the same status, being in school. Hence, the longer two sequences will be identical, the less different they should be. However, according to our results, it appears that if the purpose of the study is to build homogeneous clusters and identify the main typologies, LCS should be used instead. Thus we can assume that a trajectory is not defined by the first or the final job but rather by the proximity of the transitions during the career-path. As in the previous example, the quality indexes in Table 2 show that the use of an optimally-tuned combination of dissimilarities provides a nice trade-off between clustering (the quantization error) and mapping (the topographic error).

OM, LCS and χ^2 metrics. Finally, the MR-SOM was run with the three OM metrics, the best Elzinga dissimilarity, LCS, and the χ^2-distance. According to the results in Table 3, the χ^2-distance has the most important weight and it contributes the most to the resulting clustering. The weights of the other dissimilarities are generally below 5%. The clustering and the resulting typologies are then defined by the contemporaneity of their identical situations, rather then by the transitions or the common subsequences. Hence, it appears that the timing and not the duration or the order is important for the clustering procedure.

a) Optimally-tuned α

Metric	LCP	RLCP	LCS
α-Mean	0.02739	0.00228	0.97032
α-Std	0.02763	0.00585	0.02753

b) Quality criteria for the SOM-clustering

Metric	LCP	RLCP	LCS	Optimally-tuned α
Quantization error	379.77573	239.63652	93.50893	107.1007
Topographic error	0.07788	0.04344	0.07660	0.0495
Prototype dispersion	2693.47676	2593.21763	2094.27678	2080.8514

Table 2. Preliminary results for three Elzinga metrics

This confirms the importance of the history on the identification of a trajectory. Some temporal events are crucial on the labor market and a common behavior during these periods is determinant to define a common typology. However, let us remark two things. On the one hand, the quantization error is significantly improved, hence the clustering properties of the mixture of the five dissimilarities are better than for the previous examples. On the other hand, the topographic error becomes very large, hence the mapping properties are degraded. The combination of the five dissimilarities is then particularly adapted for extracting typologies, but is less interesting for visualization purposes.

a) Optimally-tuned α

Metric	OM	HAM	DHD	LCS	χ^2
α-Mean	0.06612	0.03515	0.03529	0.03602	0.82739
α-Std	0.04632	0.02619	0.02630	0.03150	0.07362

b) Quality criteria for the SOM-clustering

Metric	Optimally-tuned α
Quantization error	75.23233
Topographic error	0.56126
Prototype dispersion	484.00436

Table 3. Preliminary results for the five best dissimilarities

5.1 Results on the Whole Data Set

In addition to the statistical indexes computed in the previous section, we can compare different dissimilarities by inspecting the resulting self-organizing maps. Three maps were trained on the whole data set : the

first is based on the χ^2-distance, the second on the best performing Elzinga metric in the above section, the length of the longest subsequence (LCS), while the third was obtained by running online multiple-relational SOM on the three optimal-matching dissimilarities (OM, Hamming, DHD). We can note that the three maps provide some common paths: a fast access to permanent contracts (clear blue), a transition through fixed-term contracts before obtaining stable ones (dark and then clear blue), a holding on precarious jobs (dark blue), a public temporary contract (dark green) or an on-call (pink) contract ending at the end by a stable one, a long period of inactivity (yellow) or unemployment (red) with a gradual return to employment. The maps obtained by LCS and OM dissimilarities are quite similar. A drawback of the OM map is its difficulty to integrate paths characterized by a long return in the educative system (purple). This path is better integrated in the LCS map. The visual interpretation of the two maps gives support to the OM map due to a progressive transition on the map between trajectories of exclusion on the west and quick integration on the east. This reading is less clear on the LCS map. The χ^2 map is a little bit different: we observe more different trajectories (by example a start by apprenticeship contract (clear green) ending with a fixed-term or a permanent-term contract). The reading of the map is easy without any outliers paths and a clear distinction of the trajectories between north (exclusion - education in west, unemployment in east), middle (specific short-term contracts - public, apprenticeship and on-call from west to east) and south (integration - long term contracts in east, short term ones in east). Overall its diversity and its ease to read give support to the χ^2 map against the LCS and OM ones. This confirms that the overweighting of the χ^2-distance on the five dissimilarities could be attributed to a better fit of this dissimilarity on our dataset.

6 Conclusion and future work

A modified version of online relational SOM, capable of handling several dissimilarity matrices while automatically optimizing a convex combination of them, was introduced. The algorithm was used for analyzing life sequences for which the question of selecting an appropriate metric is largely debated. Instead of one dissimilarity, we used several categories that were automatically mixed in an optimal combination.
As explained in the previous section, the main drawback of the proposed relational SOM algorithm is related to the computation time. We are currently investigating a sparse version which will allow us to handle very large datasets.

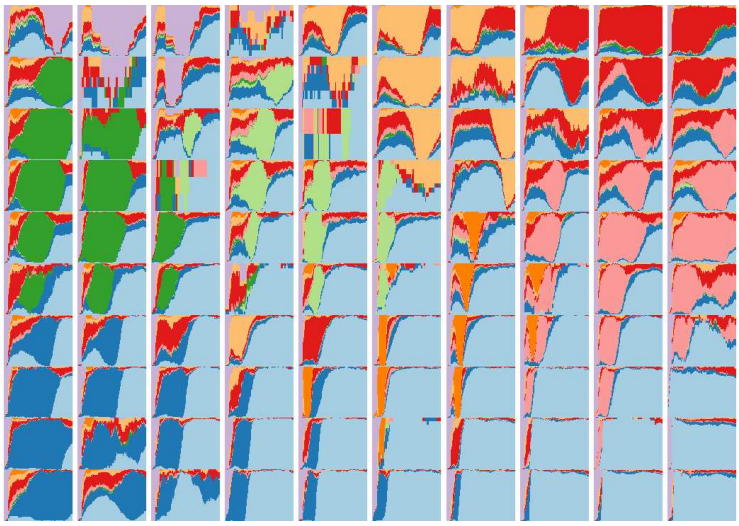

Fig. 2. Final map obtained with the χ^2-distance

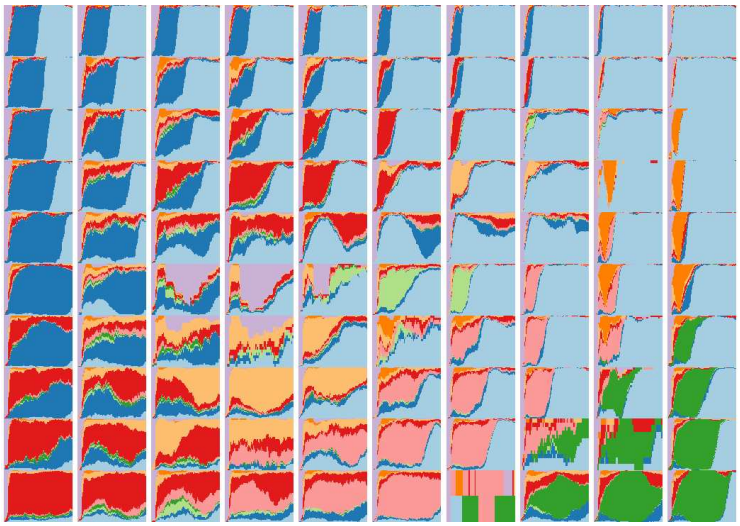

Fig. 3. Final map obtained with the LCS-dissimilarity

References

1. Elzinga, C.H.: Sequence similarity: a nonaligning technique. Sociological methods et research **32** (2003) 3–29

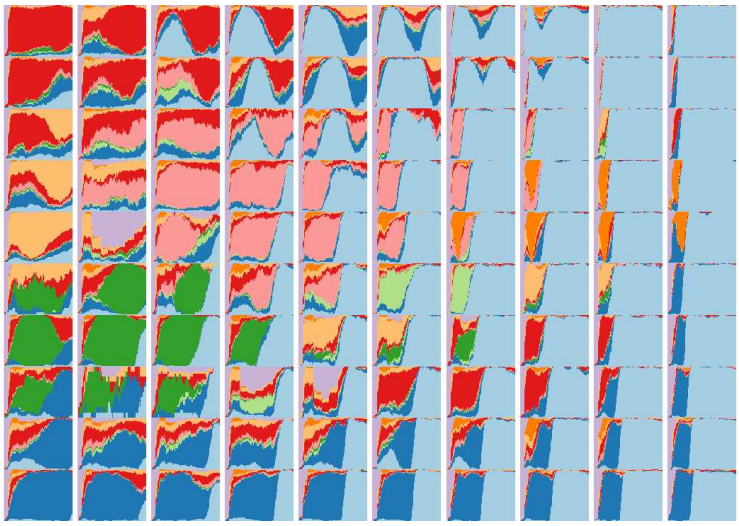

Fig. 4. Final map obtained with the OM dissimilarities

2. Robette, N.: Explorer et décrire les parcours de vie: les typologies de trajectoires. CEPED ("Les Clefs pour"), Université Paris Descartes. (2011)
3. Jean-Pierre Fénelon, Yvette Grelet, Yvette Houzel: The sequence of steps in the analysis of youth trajectories. E.J.E.S.S. **14**(1) (2000) 27–36
4. Olteanu, M., Villa-Vialaneix, N., Cottrell, M.: On-line relational som for dissimilarity data. In: Advances in Self Organizing Maps, 9th International Workshop WSOM 2012, Santiago, Chile, December 12-14, 2012, Proceedings. Volume 198 of Advances in Intelligent Systems and Computing., Springer (2012) 13–22
5. 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) (1970) 443–453
6. Abbott, A., Forrest, J.: Optimal matching methods for historical sequences. Journal of Interdisciplinary History **16** (1986) 471–494
7. Abbott, A., Tsay, A.: Sequence analysis and optimal matching methods in sociology. Review and prospect. Sociological Methods and Research **29**(1) (2000) 3–33
8. Wu, L.: Some comments on "sequence analysis and optimal matching methods in sociology, review and prospect". Sociological methods and research **29**(1) (2000) 41–64
9. Müller, N., Ritschard, G., Studer, M., Gabadinho, A.: Extracting knowledge from life courses: Clustering and visualization. In: Data Warehousing and Knowledge Discovery, 10th International Conference DaWaK 2008, Turin, Italy, September 2-5, LNCS 5182, Berlin: Springer. (2008) 176–185

10. Lesnard, L.: Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. Sociological Methods et Research **38**(3) (2010) 389–419
11. Elzinga, C.H.: Sequence analysis: metric representations of categorical time series. Sociological methods and research (2006)
12. Kohonen, T.: Self-Organizing Maps, 3rd Edition. Volume 30. Springer, Berlin, Heidelberg, New York (2001)
13. Olteanu, M., Massoni, S., Rousset, P.: Career-path analysis using optimal matching and self-organizing maps. In: José Carlos Príncipe, Risto Miikkulainen (Eds.): Advances in Self-Organizing Maps, 7th International Workshop, WSOM 2009, St. Augustine, FL, USA, June 8-10, 2009. Proceedings. Lecture Notes in Computer Science 5629 Springer 2009, ISBN (2009) 154–162
14. Conan-Guez, B., Rossi, F., El Golli, A.: Fast algorithm and implementation of dissimilarity self-organizing maps. Neural Networks **19**(6-7) (2006) 855–863
15. Mac Donald, D., Fyfe, C.: The kernel self organising map. In: Proceedings of 4th International Conference on knowledge-based intelligence engineering systems and applied technologies. (2000) 317–320
16. Hammer, B., Hasenfuss, A., annd Strickert M. Rossi, F.: Topographic processing of relational data. In: Proceedings of the 6th Workshop on Self-Organizing Maps (WSOM 07), Bielefeld, Germany (September 2007) To be published.
17. Olteanu, M., Villa-Vialaneix, N., Cierco-Ayrolles, C.: Multiple kernel self-organizing maps. (2013)
18. Rakotomamonjy, A., Bach, F.R., Canu, S., Grandvalet, Y.: Simplemkl (2008)
19. R Development Core Team: R: A Language and Environment for Statistical Computing, Vienna, Austria (2012) ISBN 3-900051-07-0.
20. Gabadinho, A., Ritschard, G., Müller, N., Studer, M.: Analyzing and visualizing state sequences in r with traminer. Journal of Statistical Software **40**(4) (2011) 1–37
21. Pölzlbauer, G.: Survey and comparison of quality measures for self-organizing maps. Volume Proceedings of the Fifth Workshop on Data Analysis (WDA'04)., Elfa Academic Press (2004) 67–82