DEMOGRAPHIC RESEARCH

A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 51, ARTICLE 16, PAGES 553–576
PUBLISHED 3 SEPTEMBER 2024
https://www.demographic-research.org/Volumes/Vol51/16
DOI: 10.4054/DemRes.2024.51.16

Research Material

Tools for analysing fuzzy clusters of sequences data

Raffaella Piccarreta

Emanuela Struffolino

Contents

Demographic Research: Volume 51, Article 16
Research Material

Tools for analysing fuzzy clusters of sequences data

Raffaella Piccarreta[1]

Emanuela Struffolino[2]

Abstract

BACKGROUND
Sequence analysis is a set of tools increasingly used in demography and other social sciences to analyse longitudinal categorical data. Typically, single (e.g., education trajectories) or multiple parallel temporal processes (e.g., work and family) are analysed by using crisp clustering algorithms that reduce complexity by partitioning cases into exhaustive and mutually exclusive groups. Crisp partitions can be problematic when clusters are not clearly separated, as is often the case in social-science applications. An effective alternative strategy is fuzzy clustering, allowing cases to belong to different clusters with a different degree of membership.

OBJECTIVE
We extend the scarce literature on fuzzy clustering of sequences to the analysis of multiple trajectories jointly unfolding over time. We illustrate how to properly apply fuzzy algorithms in this case. We propose some criteria (the fuzzy silhouette coefficients) to support the choice of the number of clusters to extract, and we introduce the gradient index plot to enhance the substantive interpretation of (multichannel) fuzzy-clustering results.

METHODS
We first describe the general features of fuzzy clustering applied to sequence data. We then use an illustrative example of multidomain sequence analysis applied to family and work trajectories to present the fuzzy silhouette coefficient and the gradient index plot.

CONTRIBUTION
These research materials provide practitioners with analytical and graphical tools that facilitate the use of fuzzy-clustering algorithms to address research questions concerning

[1] Department of Decision Sciences, Carlo F. Dondena Centre for Research on Social Dynamics, Bocconi University, and BIDSA (Bocconi Institute for Data Science and Analytics), Milano, Italy. https://orcid.org/0000-0002-8876-3656 Email: raffaella.piccarreta@unibocconi.it.

[2] Corresponding author. Department of Social and Political Sciences, University of Milan, Milano, Italy. https://orcid.org/0000-0002-6635-8748. Email: emanuela.struffolino@unimi.it.

the link between the unfolding of multiple trajectories in sequence analysis, for demographic research and beyond.

1. Introduction

Sequence analysis (SA henceforth; see Raab and Struffolino (2022) for an introduction) is a set of tools increasingly used in demography and social sciences to analyse life courses and, more in general, longitudinal data. In its standard formulation, SA is used to describe single temporal processes, such as employment (Devillanova, Raitano, and Struffolino 2019), family formation (Di Giulio, Impicciatore, and Sironi 2019), or housing trajectories (Mikolai and Kulu 2019). More recently, increasing attention has been devoted in the literature to the use of SA to describe parallel processes – for example, work and family trajectories (Aisenbrey and Fasang 2017; Rowold, Struffolino, and Fasang 2023) or partnership, sexual behaviour, and contraceptive trajectories (Brew et al. 2020). Trajectories are operationalised as sequences of the *states* experienced by individuals over a specific period, observed at regular time points (month, semester, or year).[3]

Most of the applications of SA heavily rely on cluster analysis: Because individual trajectories differ one from another to a varying extent (and the more so when several trajectories are considered in combination), some sort of data reduction is necessary. Cluster analysis allows researchers to partition cases into (as much as possible) homogeneous groups. Such partition is interpreted as a typology whose types are regarded as the typical patterns of the social process under analysis as it unfolds over time. Given a matrix defining the pairwise dissimilarity between individuals' trajectories,[4] almost all the contributions in the literature obtain clusters by using the Ward's hierarchical algorithm or the partitioning around medoids (Kaufman and Rousseeuw 1990) algorithm (PAM). These crisp (or hard) algorithms lead to an exhaustive partition of mutually exclusive clusters. Besides being used to ease the description of the most relevant tendencies in data, in many applications, the multinomial variable indicating the cases' cluster membership is used as a dependent or as an independent variable in linear and nonlinear regression models.

[3] For example, the family-formation trajectory of an individual might describe her partnership status in each time point (e.g., single, in a union, married, or divorced), her parenthood status (e.g., childless, with one child, or with two children), or a combination of the two dimensions.

[4] Many proposals have been introduced in the literature to assess the dissimilarity between sequences properly, depending on the trajectories' features, which – according to the researcher – should be mostly focused on in such evaluation. We do not account for the different available options here and refer interested readers to the extended review by Studer and Ritschard (2016).

Some authors (e.g., Liao et al. 2022; Piccarreta and Studer 2019) underline that crisp clustering algorithms forcedly assign all the cases to one and only one cluster, even if some sequences deviate from their cluster and have little in common with the other cases. Ignoring the within-cluster heterogeneity due to the potential differences among cases (especially when the number of clusters is not particularly high) might lead to an overly simplistic description of the data structure or to wrong or inaccurate conclusions when clusters are used in regression models.

To date, only a few proposals have been advanced in the SA literature to account for the uncertainty of the allocation of cases to clusters: Jalovaara and Fasang (2020) rely on the silhouette coefficients (Rousseeuw 1987) to identify cases poorly related to clusters. Piccarreta and Struffolino (2024) underline that even if silhouette coefficients allow for the identification of sequences that lie between clusters, they fail to identify cases that are relatively distant from their cluster yet are not close to any other cluster. Piccarreta and Struffolino (2024) furthermore introduce alternative criteria to inspect clusters' internal composition and to detect cases characterised by rare patterns or outliers that compromise cluster homogeneity.

In a recent contribution, Helske, Helske, and Chihaya (2023: 3) point out that "in social sciences it is unrealistic to assume that any true underlying clusters exist." In addition, even if true clusters exist, it might be difficult to identify them (as emphasised – among the others – by Warren et al. [2015]) because different partitions might be identified by different algorithms. Moreover, Helske, Helske, and Chihaya (2023) highlight that this issue might be particularly serious in SA applications since different pairwise dissimilarities measures can be used. Thus, it is important to account for the fact that the assignment of cases to clusters might be imperfect by properly addressing the uncertainty of the clustering results.

In this regard, Studer (2018) suggests employing soft rather than crisp (or hard) algorithms to cluster sequence data. Instead of assigning each case to one cluster only, soft algorithms allow cases to belong to more than one cluster and identify a different degree of membership to each of the clusters. In particular, Studer (2018) refers to fuzzy-clustering algorithms that – differently from other soft algorithms – can be applied also when only dissimilarities (and no measurements on a set of variables) are available, as is the case in SA applications (see, nonetheless, Murphy et al. (2021) for an example of model-based clustering applied to very particular – though seldomly used in SA – dissimilarities based on the matches between the activities experienced in each time period). Helske, Helske, and Chihaya (2023) make a detailed exploration of the advantages of fuzzy over crisp clustering, and are able to show that the former is recommended when the cluster structure is weak because it allows for a better description of the multifaceted characteristics of cases lying at the borders or between multiple clusters and of the peculiar features of cases that are distant from all the clusters. Even if

Helske, Helske, and Chihaya (2023) focus specifically on the robustness of results obtained when clusters are used as outcomes or predictors in regression models, their considerations have more general implications, because – as mentioned above – cluster analysis is primarily used in applications to simplify and explore the sequences' structure.

Despite its touted advantages, fuzzy clustering is not particularly popular in SA applications: So far, it has been used solely for the analysis of single trajectories (e.g., Salem et al. 2016; Studer 2018). A possible reason for this is that, compared to crisp clusters, fuzzy clusters might be more difficult to interpret and visualise. This is probably why, to the best of our knowledge, there are no contributions in the literature that use fuzzy-clustering algorithms in the more complex situation of analysing multiple trajectories jointly using the so-called multichannel sequence analysis (MCSA) (Pollock 2007; Gauthier et al. 2010; Piccarreta 2017; Ritschard, Liao, and Struffolino 2023). MCSA is receiving increasing attention in empirical applications. In this type of analysis, however, partitioning cases is particularly important – also to simply explore the data and to get some understanding about how multiple trajectories evolve together. In addition, in the case of MCSA the issues of within-cluster homogeneity and of the potential misallocation of sequences when the clustering structure is weak becomes even more serious than in single-channel applications.

Indeed, cases characterised by uncommon (if not rare) combinations of trajectories are frequent, and defining a typology unveiling all the relevant combinations of traits in the involved domains can be more challenging because of the inherent higher complexity of the data. Since clustering algorithms prioritise the identification of the most common patterns, peculiar cases or marginal groups of cases might not be placed in dedicated clusters, even when a large number of clusters is extracted. Typically, these cases lie at the edges of different clusters, being similar to one cluster with respect to the trajectories in one domain and to (one or more of the) others with respect to the trajectories in other domains, or are distant from all clusters.

We argue that fuzzy clustering can prove particularly beneficial for MCSA. By allowing cases to belong to more than one cluster, it permits one to focus on the more relevant combinations of trajectories, thus maintaining a reasonably low number of clusters and easing the substantive interpretation of the results. It also allows for the easy identification and characterisation of cases that are isolated or weakly related to the obtained clusters. Consistently, the goal of the research materials in this contribution is to illustrate how to properly apply fuzzy algorithms for the analysis of multichannel sequences so as to foster a more conscious and intensive use of fuzzy clustering in SA applications. Even if our attention is focused on the more complex case of multiple trajectories, the described procedures can be useful to analyse fuzzy clusters of single trajectories as well.

The following materials are organised as follows. We first describe the general features of the most commonly used fuzzy algorithm. Second, we review extensions to fuzzy clustering of the average silhouette coefficient (Kaufman and Rousseeuw 1990) used to assess and/or monitor the quality of alternative crisp partitions. We then apply the algorithm to cluster a sample of Italian women based on the work- and family-related events they experienced between the ages of 18 and 35. Finally, we introduce the gradient index plot, a novel visualisation tool to enhance the substantive interpretation of (multichannel) fuzzy clusters. All in all, we are hereby providing practitioners with analytical and graphical tools that will facilitate the use of fuzzy-clustering algorithms in SA applications (in demographic research and beyond) to address research questions concerning single as well as multiple trajectories that unfold jointly over time.

2. Fuzzy clustering

For a given number of clusters C, a fuzzy algorithm (Kaufman and Rousseeuw 1990; Maechler et al. 2005) assigns to each case, say the i-th, a vector $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{iC})$ whose elements, ranging between 0 and 1 and summing up to 1, indicate the degree to which the case belongs to each cluster. More specifically, when the algorithm is based on distances/dissimilarities (as in the case of SA) rather than on vectors of measurements, the vectors $\boldsymbol{u}_i$, for $i = 1, \ldots n$, are determined by minimising the function

$$\sum_{c=1}^{C} \frac{\sum_{i,h=1}^{n} u_{ic}^{r} u_{hc}^{r} \delta_{ih}}{2\sum_{i=1}^{n} u_{ic}^{r}}, \tag{1}$$

where δ_{ih} is the dissimilarity between two cases (assessed in our case using MCSA, see below for details), and $r > 1$ is a user-defined parameter that controls the fuzziness of the clusters.

A value of r close to 1 leads to crisper partitions (with one membership grade close to 1 and all the others close to 0), whereas higher values of r increase the level of fuzziness. Note that in the case of complete fuzziness – that is, when $u_{ic} = (1/C)$ for each case and each cluster – the solution is meaningless because each case would belong to every cluster to the same extent, and therefore clusters would not be distinguishable one from another. Similar considerations hold in the case of partially complete fuzziness, occurring when two or more clusters are characterised by the same exact membership coefficients. For the sake of illustration, Table 1 displays the membership degrees of the first four fictive cases in a five-cluster partition, where all the cases have the same degrees of membership to the last three clusters.

Table 1: Example of the fuzzy-cluster memberships for the first four cases in five clusters with partially complete fuzziness

Case	Clusters' membership				
	1	2	3	4	5
1	0.10	0.09	**0.27**	**0.27**	**0.27**
2	0.56	0.15	**0.09**	**0.09**	**0.09**
3	0.14	0.46	**0.13**	**0.13**	**0.13**
4	0.09	0.14	**0.26**	**0.26**	**0.26**
…					

Note: Bold is used to indicate values of cluster membership that are the same across clusters for the same case, that is with partially complete fuzziness.

Clearly, clusters 3 to 5 in Table 1 cannot be distinguished one from another. This situation typically occurs when either the number of clusters is too high, when the algorithm converges to a suboptimal solution, or when the parameter r is set too high. The optimal choice for the value of r is an open issue. The value recommended in the literature is 2, even if some studies have shown that the proper choice depends highly on the data (Abadpour 2016; Yu, Cheng, and Huang 2004), and therefore different values of the parameter should be considered (see also Pal and Bezdek 1995 and Zhou, Fu, and Yang 2014). A common data-driven procedure to select r is to start setting $r = 2$ and reduce r in case of non-convergence or of complete or partially complete fuzziness.

3. Choosing the number of clusters: The fuzzy silhouette coefficient

A relevant issue in cluster analysis concerns the choice of the number of clusters. Whilst a final decision about the number of clusters should (also) be based on the substantive interpretation of the clusters and on the patterns identified in data (particularly when clusters are employed in regression analyses), it is recommendable to monitor the quality of partitions for varying number of clusters (see also Studer [2021]). Some authors (e.g., Kaufman and Rousseeuw [1990]) propose to monitor fuzzy partitions relying on standard measures of adequacy calculated on the crisp clusters derived by hardening the fuzzy ones and assigning cases to clusters according to the maximal membership degree. However, this approach does not take into account the most distinguished characteristics of the fuzzy partition properly – that is, the different degrees of membership of cases to clusters.

To address this issue, some proposals have been introduced to define suitable fuzzy counterparts of the well-known silhouette coefficient, introduced by Kaufman and

Rousseeuw (1990) for crisp partitions. In its original formulation, for the i-th case the coefficient is defined as

$$s_i = \frac{b_i - a_i}{\max(b_i, a_i)}, \tag{2}$$

where a_i is the average distance of the case from the other cases placed into its cluster, and b_i is the minimum average distance of the case from those in the other clusters. A value of s_i close to 1 indicates that the case is clustered appropriately. If instead s_i is close to -1, the case should be placed in another cluster (the closest one). A silhouette coefficient close to 0 indicates that the case lies between two clusters. The global quality of a partition can be assessed by referring to the average of the cases' silhouette coefficients.

One possible extension of the coefficient to fuzzy partitions is the so-called density-based silhouette coefficient (originally designed for model-based clusters; Menardi [2011]), that accounts for the membership degree. For the i-th case the index is defined as

$$ds_i = \frac{\log(u_{i*}^r / u_{i**}^r)}{\max_{h=1,\dots,n}[\log(u_{h*}^r / u_{h**}^r)]}, \tag{3}$$

where u_{i*} and u_{i**} are the largest and the next largest membership degrees for the case. Thus, an asterisk (*) indicates the cluster the i-th case would be assigned to based on the highest membership degree, and a double asterisk (**) indicates the cluster (other than *) the case is closest to (again, based on the membership degree). Therefore, ds_i is the ratio between the relative strength of cluster membership for the i-th case and the maximum relative strength observed in data. The fuzzy average silhouette coefficient is defined as the average of the ds_i. Note that ds_i does not depend on the dissimilarities between one case and the others, and that it does not coincide with or tend to the crisp coefficient s_i in (1) for crispier partitions – that is, when each case has a degree of membership close to 1 with one cluster and close to 0 with all the others.

Rawashdeh and Ralescu (2012) review a number of fuzzy-cluster quality indices and introduce an interesting fuzzy generalisation of the silhouette coefficient based both on the memberships and on the dissimilarities. Specifically, for each pair of cases, say the i-th and the h-th, the authors consider a measure of intra-distance within a given cluster, say the c-th

$$intra_c(i, h) = \min(u_{ic}, u_{hc}), \tag{4}$$

and a measure of inter-distance between two clusters, say the c-th and the k-th

$$inter_{c,k}(i,h) = \max[\min(u_{ic}, u_{hk}), \min(u_{ik}, u_{hc})]. \tag{5}$$

The compactness and the separations distances, a_i and b_i at the basis of the silhouette coefficients are defined as

$$a_i = \min_{c}\left[\frac{\sum_{h=1,h\neq i}^{n} intra_c(i,h)\delta_{ih}}{\sum_{h=1,h\neq i}^{n} intra_c(i,h)}\right], \text{ and} \tag{6}$$

$$b_i = \min_{c,k}\left[\frac{\sum_{h=1,h\neq i}^{n} inter_{c,k}(i,h)\delta_{ih}}{\sum_{h=1,h\neq i}^{n} inter_{c,k}(i,h)}\right]. \tag{7}$$

The generalised silhouette coefficient, gs_i, which is based on measures (6) and (7), coincides with the crisp silhouette coefficient in (1) when the partition is crisp. The dependency of such an index on dissimilarities is particularly useful in the case of MCSA. Indeed, as underlined by Piccarreta (2017), it is important to assess whether the clusters based on multichannel dissimilarities, which combine information on different domains, are homogeneous across all the considered domains, or if some domains prevail over the others, thus driving the clustering process. The quantities a_i and b_i in (6) and (7) can be evaluated using both the multichannel dissimilarity and the dissimilarities defined on each domain separately, in order to assess the quality of the fuzzy partition both at the joint level and the marginal level.

Note that – as mentioned before – silhouette coefficients offer some insights about the level of compactness of partitions and are useful to identify some plausible partition. Even so, it is crucial to analyse and compare alternative clusters' composition in order to select a partition which is meaningful from the substantive point of view. For this reason, it is important to offer some criteria to properly describe the fuzzy clusters' most salient features. We propose a novel graphical tool to do so. For the sake of clarity and exposition, we will introduce it in the following section framed by a discussion of fuzzy-clustering application to analyse a sample of Italian women's work and family trajectories.

4. Illustrative application of fuzzy-clustering to employment–family sequences

4.1 Data

We used data from the Multi-purpose Survey on Household and Social Subjects carried out in 2009 by the Italian National Statistical Office. The survey collected retrospective information on all household members on several life domains, notably educational, employment, and family episodes. The data consist of a representative sample of around 43,850 Italian residents. Here we focus on 6,801 women born between 1955 and 1975, whose employment and family-formation activities could be observed between the ages 18 and 35.

For each woman in the sample, we reconstructed the monthly sequences describing the work- and family-related events experienced between the ages of 18 and 35. To code the employment trajectories, we distinguished between education, out-of-labour force (which includes unemployment and inactivity), and part-time and full-time work.[5] To code family-formation trajectories, we focused on three partnership statuses – being single, in a union (cohabitation or marriage), or divorced (or separated) – and added a specification for the number of children depending on the partnership status. For women in a union, we distinguished between having no children, only one child, or two or more children (0, 1, 2+), whereas we distinguished single or divorced women only as either having children or not (0, 1+). This latter decision was motivated by the fact that the number of women who had children out of a union or who divorced before turning 35 was relatively low, and we wanted to avoid using states that were too rare.[6]

[5] Part-time and full-time work included dependent permanent employment, temporary employment, and self-employment, depending on the working hours of each episode in these contractual arrangements. Part-time and full-time work also included months spent in parental leave when a woman was working either full-time or part-time before the leave.

[6] All the elaborations were obtained using the software R version 4.3.1 (R Core Team 2022). Sequences were built using packages *TraMineR* (Gabadinho et al. 2011) and *WeightedClusters* (Studer 2013). Fuzzy clusters were extracted using the function *fanny* in package *cluster* (Maechler et al. 2005). The code to calculate the density-based and the generalised silhouette coefficients and to build the weighted index sequence plots in Figure 2 are available as additional files on the journal webpage, as well as on the GitHub page of one of the authors, https://github.com/raffaellapiccarreta/Tools-for-analysing-fuzzy-clusters-of-sequences-data.

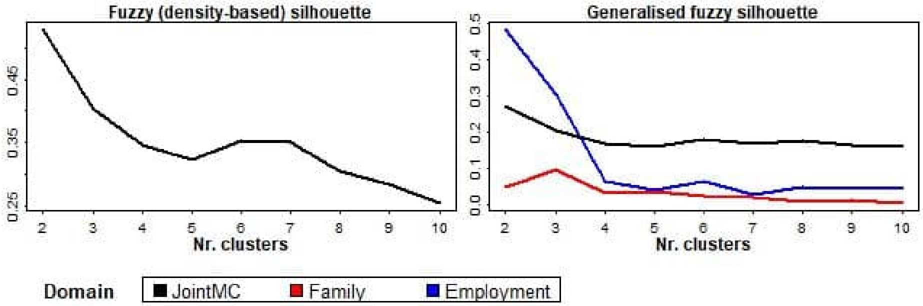
4.2 Fuzzy-clustering employment-family sequences

In a first preliminary step, we computed the dissimilarities between women in the sample based on the combination of work and family trajectories using MCSA. MCSA relies on ‘edit’ dissimilarities, which assess the difference between two sequences (in one domain) by quantifying the total cost of the operations needed to transform one sequence into the other. Following the approach proposed by Pollock (2007), we referred to the edit dissimilarity based on the very popular optimal matching algorithm (Abbott 1995), which focuses on three basic operations: insertion or deletion of a state and substitution of one state for another. MCSA combines the (edit) dissimilarities for each domain into one multichannel dissimilarity by averaging the costs of the operations needed to align the sequences in each domain. Doing so preserves the information on each domain as measured by specific costs. We applied the fuzzy algorithm developed by Kaufmann and Rousseeuw (1990) for dissimilarity matrices, and extracted a number of clusters between 2 and 10. As discussed in Section 2, to tune the fuzzifier parameter r, we started with a value of $r = 2$, and progressively decreased r in case of no convergence or of complete or partial fuzziness. The first value of r leading to converging solutions with suitable memberships for each number of clusters was $r = 1.3$.

To monitor fuzzy partitions with different numbers of clusters, we considered (Figure 1) both the density-based average silhouette coefficients (Menardi 2011) and the generalised average silhouette coefficients (Rawashdeh and Ralescu 2012) illustrated in Section 3. The former approach, based solely on the relative strength of clusters’ memberships points to partitions with 6 or 7 clusters, if one excludes the 2-clusters solution, which is possibly too simplistic.

To assess both the joint and the marginal performances of the obtained clusters, the generalised average silhouette coefficients were calculated based on the multichannel dissimilarities and on the optimal matching dissimilarities built for each domain separately. At the joint-level, the coefficient remained quite stable for partitions with 6 clusters or more and showed a slight decline when moving from 6 to 5 clusters. Thus, the solutions with 6 or 7 clusters appeared again to be reasonable compromises. At the domains-specific level, the silhouette coefficients were quite well aligned, even if clusters appeared (slightly) relatively more compact and distinguished for the employment domain (in blue), at least when the number of clusters was 6 or more.

Figure 1: Monitoring fuzzy partitions with a different number of clusters using the average density-based and the average generalised silhouette coefficients



Source: Multi-purpose Survey on Household and Social Subjects 2009.
Notes: The former coefficients are based on only the relative strength of membership within each cluster. The latter coefficients depend on both the cases' membership degrees and their dissimilarities, and were obtained referring both to the multichannel dissimilarities (JointMC) used to build clusters and to the domains-specific dissimilarities obtained separately for each domain (Family and Employment). This allows one to assess the ability of the clusters to account also for dissimilarities in each specific domain.

For a final decision about the number of clusters, it is recommendable – if not strictly necessary – to explore substantively and to compare the partitions identified as plausible by the considered criteria. To this aim, in the next section we will illustrate our proposal for an effective visualisation of fuzzy clusters of (multiple) sequences.

5. Visualising fuzzy clusters of sequences: The gradient index plot

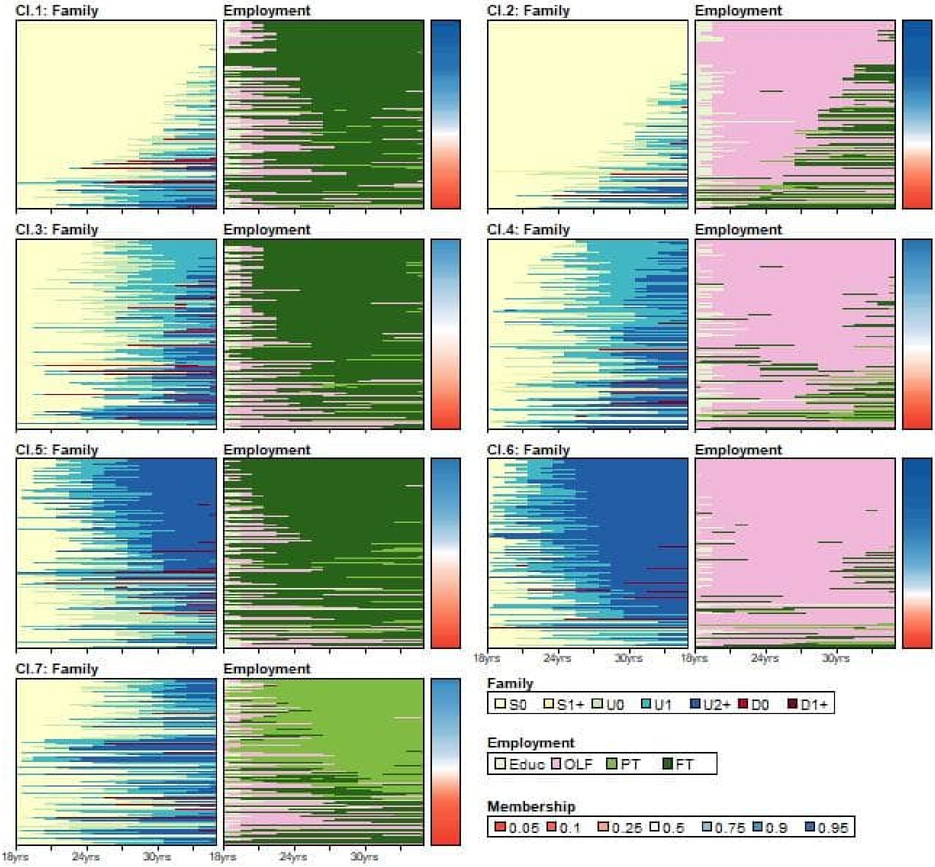
Here, we introduce novel graphical tools to suitably visualise fuzzy clusters of sequences and to compare partitions focusing on more substantive considerations based on the clusters' composition, and we demonstrate how we used them to select the final partition.

A standard tool to visualise a group of sequences is the index plot (Scherer 2001), where trajectories (on the vertical axis) are represented by horizontally stacked bars whose colours change depending on the state visited along the time line (the horizontal axis). For crisp partitions, cluster-specific index plots are used to represent only the sequences assigned to each cluster (Figure A-1 in the Appendix, Panel A). For fuzzy clusters, though, all the cases belong to each cluster, as what changes is the degree of membership. Focusing on the single-channel case, Studer (2018) proposes to represent each fuzzy cluster using a weighted index plot displaying all the sequences ordered on the vertical axis based on their degree of membership to the cluster, which also regulates the height of the sequences' horizontal bars. In the case of multichannel sequences, this allows one to identify the most salient characteristics of the parallel unfolding of trajectories (in the different domains) with the highest degrees of membership in each cluster (Figure A-1 in the Appendix, Panel B). Nonetheless, such plots do not provide information about the strength of the membership of the sequences to the clusters. Nor do they provide evidence of the combinations of trajectories whose degree of membership falls below a certain threshold. Indeed, when the number of cases is relatively high, it is not possible to fully appreciate the differences in the heights of the bars, and to deduce the strength of membership. To tackle this issue, Studer (2018) suggests considering and comparing the average membership values for each cluster in order to gain some insights about their 'compactness' (Figure A-1 in the Appendix, Panel C). Nonetheless, the distribution of the clusters' membership degrees is typically skewed to the left – with many cases characterised by very low membership; therefore, the mean might be a misleading summary measure in this case.

To gain insights about the whole distribution of the clusters' membership degrees, we enhance the weighted index plots with a gradient bar, whose colours reflect the level of membership for the different cases. This gradient index plot (see Figure A-1 in the Appendix, Panel D) allows one to exploit the granular information provided by fuzzy clusters and to identify the core of the clusters (i.e., cases with the highest degrees of membership) as well as those cases that are related to more clusters or that do not fully belong to any cluster. This enhances the interpretation of the clusters and the comparison of their relative strength. Indeed, note that while for crisp clusters it is possible to evaluate the average silhouette coefficient for each cluster, this is not possible for fuzzy clusters.

Figure 2 displays the gradient index plot for the 7-clusters solution, one of those supported by the silhouette coefficients (Figure 1). For each cluster, one plot for each domain (here, family and employment) is reported, together with the gradient bar, whose colours reflect the degrees of membership of cases to clusters (from dark red, indicating the lowest membership degree, to dark blue, indicating the highest membership degree). For each cluster, all of the 6,801 cases are displayed in the plots of the family and employment trajectories as well as in the gradient bar. Both the order of the cases on the vertical axes and the heights of the bars representing their sequences depend on the degree of membership. The gradient bars allow one to easily identify the core-sequences characterising (i.e., having the highest degrees of membership to) each cluster and the sequences with an intermediate level of 'attachment' to the cluster. In addition, these graphs permit the viewer to appreciate the decrease in the membership degree as the features of the sequences change. A detailed description of the core-sequences in each fuzzy cluster is offered in Figure A-2 in the Appendix, where for each cluster only sequences with a degree of membership higher than 0.5 are displayed.

Figure 2: Weighted gradient index plots for the seven fuzzy clusters of the family and employment trajectories



Source: Multi-purpose Survey on Household and Social Subjects 2009.
Notes: For each cluster the sequences in the two domains are ordered on the vertical axes based on their degree of membership to the cluster; the height of the horizontal bar representing each sequence is proportional to the degree of membership. Since the sequences' bars differ depending on the degree of membership, the plots do not report tick marks and labels on the y-axis. For each cluster, the gradient bar displays the degree of membership of each case to the cluster. Note that for each cluster all the sequences in the dataset are displayed: The differences across clusters are actually due to the different ordering of the sequences and to the different heights of the bars. Family: S0=single without children; S1+=single with one or more children; U0=in a union without children; U1=in a union with one child; U2+ = in a union with two or more children; D0 = divorced without children; D1+ = divorced with one or more children. Employment: Educ = education; OLF = out of labour force; PT = part-time work; FT = full-time work.

For the selected fuzzy partition, clusters 1, 2, and 6 include greater proportions of cases with a relatively high degree of membership (more dark blue in the gradient bar). Cluster 3 has a much weaker structure (that is, higher within-cluster heterogeneity), with only a few cases very close to the cluster. Cluster 7 presents the greatest proportion of cases with very low-degree memberships (more dark red in the gradient bar).

Moving to the substantive interpretation of the clusters, the core of cluster 1 are women whose employment sequences are characterised by a fast transition out of education that stabilises into full-time employment, associated with long-term singleness and childlessness in the family-formation domain. It can also be noted that the degree of membership decreases as the age at cohabitation and at first child decreases (and as the entry into the labour market is postponed); this makes the differences in the degrees of membership interpretable (to a certain extent).

The other two clusters, whose core cases present long-term full-time employment, are clusters 3 and 5, showing, however, lower degrees of membership. In these two cases, long-term full-time employment is associated with the transition to union and parenthood in their late-20s/early-30s (cluster 3) and, respectively, with an earlier transition to union and motherhood, with two or more children, by the age of 28 (cluster 5). Even in this case, it is possible to appreciate how the membership grade decreases as the number of children increases or decreases, respectively.

Cluster 2, 4, and 6 are characterised by the same family-formation dynamics as clusters 1, 3, and 5. In the case of the former, though, they are associated with permanent unemployment after secondary education. Cluster 2 and 6 in particular show high degrees of membership, and in the case of cluster 2 this holds also for a set of women who remain single and enter the labour market with full- or part-time job after the age of 30. This is interesting because, in the case of clusters 4 and 6, sequences with high degrees of membership are characterised by family-formation trajectories that seem to imply no chances to enter paid work.

Finally, the core of cluster 7 is composed by a relatively small number of women with employment trajectories characterised by long-term part-time work coupled with heterogeneous family-formation trajectories, where singleness and union are followed by parenthood between their late-20s and the mid-30s.

A comparison between the 7-cluster, the 6-cluster, and the 8-cluster partitions by using the gradient index plots can be found in Figures A-3 and A-4 in the Appendix. Reducing the number of clusters to 6 seemed not particularly recommendable because the parallel combinations of family and work trajectories identified in the 7-cluster solution and the dedication of one cluster (the 7th) to women working part-time seemed all worthwhile of identification (specifically, moving from 7 to 6 clusters, substantially the 4th cluster was removed). On the other hand, increasing the number of clusters from 7 to 8 led to the identification of an additional cluster (the 7th cluster in Figure A-4 in the Appendix) whose core is constituted by a limited number of women who entered the labour market after a period of unemployment and generally after the birth of their first child. Quite interestingly, in the 7-cluster partition, these women were characterised by medium levels of degree membership, given their family trajectories, both with the clusters characterised by long-term employment and with the clusters characterised by long-term unemployment. Therefore, one can easily identify and characterise such women by focusing on cases that do not belong to a specific cluster in the 7-cluster partition to a reasonably high extent.

These substantive considerations on the comparison between different cluster solutions demonstrate the advantage of fuzzy over crisp clusters, because cases that would be ‘misplaced’ by forcedly assigning them to one crisp cluster only are instead characterised by medium-level membership with more fuzzy clusters.

6. Conclusions

These research materials enrich the scarce literature on fuzzy clustering of sequences – limited so far to the case of single trajectories – to the analysis of multiple trajectories that jointly unfold over time (MCSA). We introduced (1) fuzzy silhouette coefficients to support the choice of the number of clusters to extract and (2) gradient index plots to enhance the substantive interpretation of multichannel fuzzy-clustering results.

Our goal going forward is to stimulate and foster the use of fuzzy-clustering algorithms to address research questions concerning the link between multiple temporal processes, as well as to analyse single processes. As remarked upon by various contributions in the literature, assumptions in social sciences that ‘true’ clusters exist and/or are well separated is generally too far-fetched. Accounting for heterogeneity and for the presence of cases sharing traits with more types explicitly is therefore crucial in order to properly describe the multifaceted characteristics of social processes.

7. Acknowledgements

The authors thank the reviewers for their useful comments and suggestions.

References

Abadpour, A. (2016). *Notes of fuzzy clustering*. https://abadpour.com/files/pdf/Jeannie.pdf.

Abbott, A. (1995). Sequence analysis: New methods for old ideas. *Annual Review of Sociology* 21(1): 93–113. doi:10.1146/annurev.soc.21.1.93.

Aisenbrey, S. and Fasang, A. (2017). The interplay of work and family trajectories over the life course: Germany and the United States in comparison. *American Journal of Sociology* 122(5): 1448–1484. doi:10.1086/691128.

Brew, B., Weitzman, A., Musick, K., and Kusunoki, Y. (2020). Young women's joint relationship, sex, and contraceptive trajectories. *Demographic Research* 42(34): 933–984. doi:10.4054/DemRes.2020.42.34.

Devillanova, C., Raitano, M., and Struffolino, E. (2019). Longitudinal employment trajectories and health in middle life: Insights from linked administrative and survey data. *Demographic Research* 40(47): 1375–1412. doi:10.4054/DemRes.2019.40.47.

Di Giulio, P., Impicciatore, R., and Sironi, M. (2019). The changing pattern of cohabitation. *Demographic Research* 40(42): 1211–1248. doi:10.4054/DemRes.2019.40.42.

Gabadinho, A., Ritschard, G., Studer, M., and Müller, N.S. (2011). *Mining sequence data in R with the TraMineR package: A user's guide*. Geneva: University of Geneva.

Gauthier, J.-A., Widmer, E.D., Bucher, P., and Notredame, C. (2010). Multichannel sequence analysis applied to social science data. *Sociological Methodology* 40(1). doi:10.1111/j.1467-9531.2010.01227.x.

Helske, S., Helske, J., and Chihaya, G.K. (2023). From sequences to variables: Rethinking the relationship between sequences and outcomes. *Sociological Methodology* 54(1). doi:10.1177/00811750231177026.

Jalovaara, M. and Fasang, A.E. (2020). Family life courses, gender, and mid-life earnings. *European Sociological Review* 36(2): 159–178. doi:10.1093/esr/jcz057.

Kaufman, L. and Rousseeuw, P. (1990). *Finding groups in data: an introduction to cluster analysis*. Hoboken, NJ: John Wiley & Sons. doi:10.1002/9780470316801.

Liao, T.F., Bolano, D., Brzinsky-Fay, C., Cornwell, B., Fasang, A.E., Helske, S., Piccarreta, R., Raab, M., Ritschard, G., and Struffolino, E. (2022). Sequence analysis: Its past, present, and future. *Social Science Research* 107(102772). doi:10.1016/j.ssresearch.2022.102772.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2005). Cluster analysis basics and extensions. R package version 2.1.6. https://CRAN.R-project.org/package=cluster.

Menardi, G. (2011). Density-based silhouette diagnostics for clustering methods. *Statistics and Computing* 21: 295–308. doi:10.1007/s11222-010-9169-0.

Mikolai, J. and Kulu, H. (2019). Union dissolution and housing trajectories in Britain. *Demographic Research* 41(7): 161–196. doi:10.4054/DemRes.2019.41.7.

Murphy, K., Murphy, T.B., Piccarreta, R., and Gormley, I.C. (2021). Clustering longitudinal life-course sequences using mixtures of exponential-distance models. *Journal of the Royal Statistical Society Series A: Statistics in Society* 184(4): 1414–1451. doi:10.1111/rssa.12712.

Pal, N.R. and Bezdek, J.C. (1995). On cluster validity for the fuzzy c-means model. *IEEE Transactions on Fuzzy Systems* 3(3): 370–379. doi:10.1109/91.413225.

Piccarreta, R. (2017). Joint sequence analysis: Association and clustering. *Sociological Methods and Research* 46(2): 252–287. doi:10.1177/0049124115591013.

Piccarreta, R. and Struffolino, E. (2024). Identifying and qualifying deviant cases in clusters of sequences: The why and the how. *European Journal of Population* 40(1). doi:10.1007/s10680-023-09682-3.

Piccarreta, R. and Studer, M. (2019). Holistic analysis of the life course: Methodological challenges and new perspectives. *Advances in Life Course Research* 41: 100251. doi:10.1016/j.alcr.2018.10.004.

Pollock, G. (2007). Holistic trajectories: A study of combined employment, housing and family careers by using multiple-sequence analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170(1): 167–183. doi:10.1111/j.1467-985X.2006.00450.x.

R Core Team (2022). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org.

Raab, M. and Struffolino, E. (2022). *Sequence analysis*. Thousand Oaks, CA: Sage. doi:10.4135/9781071938942.

Rawashdeh, M. and Ralescu, A. (2012). Crisp and fuzzy cluster validity: Generalized intra-inter silhouette index. *2012 Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS)*. Berkeley, CA: IEEE. doi:10.1109/NAFIPS.2012.6290969.

Ritschard, G., Liao, T.F., and Struffolino, E. (2023). Strategies for multidomain sequence analysis in social research. *Sociological Methodology* 53(2). doi:10.1177/00811750231163833.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20: 53–65. doi:10.1016/0377-0427(87)90125-7.

Rowold, C., Struffolino, E., and Fasang, A. (2023). Life-course-sensitive analysis of group inequalities in old age: Combining sequence analysis with the Kitagawa–Oaxaca–Blinder decomposition. *Sociological Methods and Research* online first. doi:10.1177/00491241231224226.

Salem, L., Crocker, A.G., Charette, Y., Earls, C.M., Nicholls, T.L., and Seto, M.C. (2016). Housing trajectories of forensic psychiatric patients. *Behavioral Sciences and the Law* 34(2–3): 352–365. doi:10.1002/bsl.2223.

Scherer, S. (2001). Early career patterns: A comparison of Great Britain and West Germany. *European Sociological Review* 17(2): 119–144. doi:10.1093/esr/17.2.119.

Studer, M. (2013). WeightedCluster library manual: A practical guide to creating typologies of trajectories in the social sciences with R. (LIVES Working Papers, NCCR LIVES 24). Lausanne: LIVES. doi:10.12682/lives.2296-1658.2013.24.

Studer, M. (2018). Divisive property-based and fuzzy clustering for sequence analysis. In: Ritschard, G. and Studer, M. (eds.). *Sequence analysis and related approaches. Innovative methods and applications*. Cham: Springer. doi:10.1007/978-3-319-95420-2_13.

Studer, M. (2021). Validating sequence analysis typologies using parametric bootstrap. *Sociological Methodology* 51(2): 290–318. doi:10.1177/00811750211014232.

Studer, M. and Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179(2): 481–511. doi:10.1111/rssa.12125.

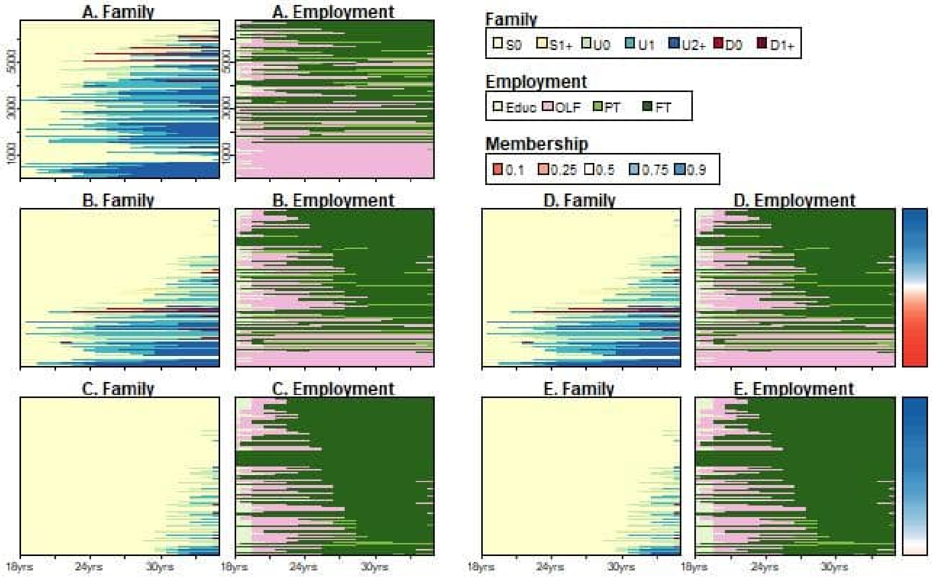
Warren, J.R., Luo, L., Halpern-Manners, A., Raymo, J.M., and Palloni, A. (2015). Do different methods for modeling age-graded trajectories yield consistent and valid results? *American Journal of Sociology* 120(6): 1809–1856. doi:10.1086/681962.

Yu, J., Cheng, Q., and Huang, H. (2004). Analysis of the weighting exponent in the FCM. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 34(1): 634–639. doi:10.1109/TSMCB.2003.810951.

Zhou, K., Fu, C., and Yang, S. (2014). Fuzziness parameter selection in fuzzy c-means: The perspective of cluster validation. *Science China Information Sciences* 57(11): 1–8. doi:10.1007/s11432-014-5146-0.

Appendix

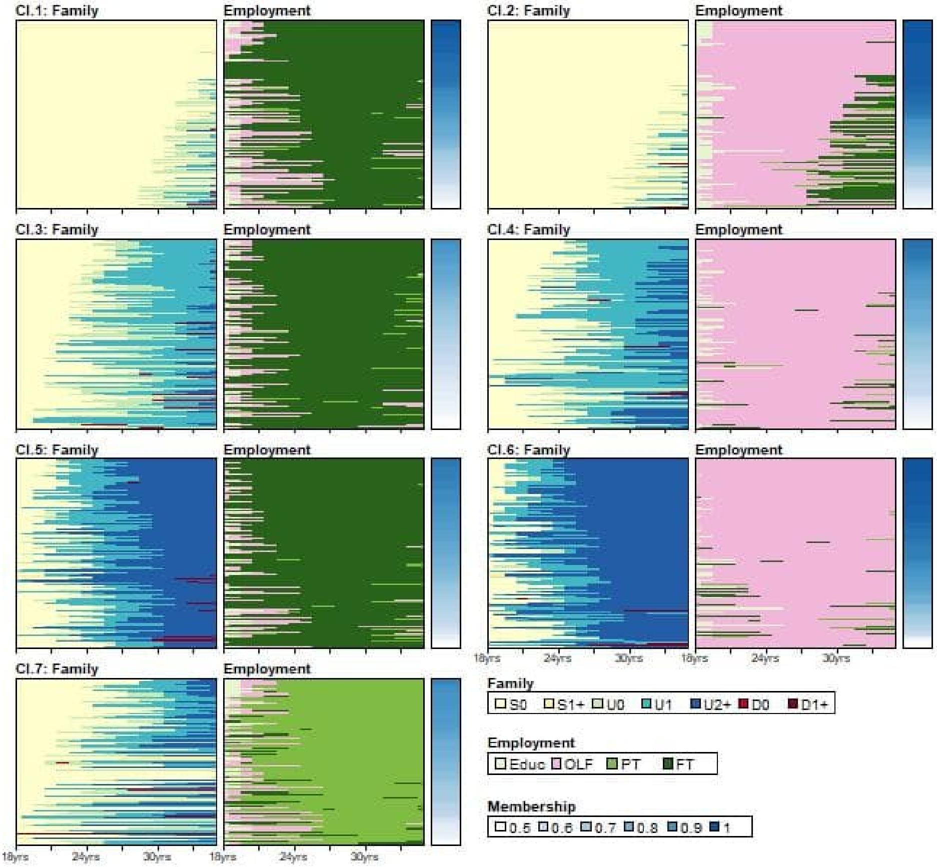
Figure A-1: Graphical display of one multichannel fuzzy cluster



Source: Multi-purpose Survey on Household and Social Subjects 2009.

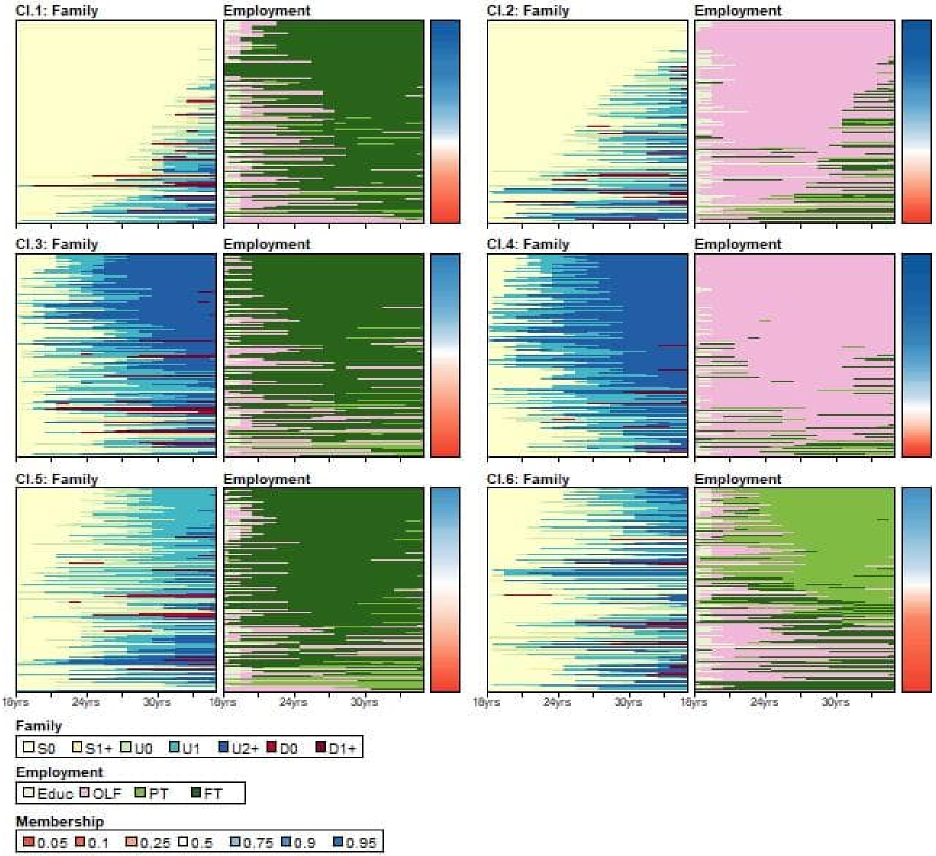
Notes: (A) Index plot reporting all the sequences in the sample, with cases ordered along the vertical axis according to the degree of membership. (B) Studer's weighted index plot: Cases are ordered on the vertical axis according to their degree of membership, and the sequences' bar heights are proportional to the degree of membership; note that also in this plot all the sequences are reported, and the differences with plot A are due to heights proportional to the degree of membership. (C) Studer's weighted index plot reporting only sequences with a degree of membership higher than 0.4. (D-E) Same as plots B-C with a gradient bar added whose colours reflect the sequences' degrees of membership to the cluster. Note that since in panels B-E the sequences' bars differ depending on the degree of membership, the plots do not report tick marks and labels on the y-axis. Family: S0 = single without children; S1+ = single with one or more children; U0 = in a union without children; U1 = in a union with one child; U2+ = in a union with two or more children; D0 = divorced without children; D1+ = divorced with one or more children. Employment: Educ = education; OLF = out of labour force; PT = part-time work; FT = full-time work.

Figure A-2: Weighted gradient index plots for the seven fuzzy clusters of the family and employment trajectories



Source: Multi-purpose Survey on Household and Social Subjects 2009.
Notes: For each cluster, (i) attention is limited to cases with a degree of membership with the cluster higher than 0.5; (ii) sequences in the two domains are ordered on the vertical axes based on their degree of membership to the cluster; (iii) the height of the horizontal bar representing each sequence is proportional to the degree of membership; and (iv) the cluster the gradient bar displays the degree of membership of each case to the cluster. Because the sequences' bars differ depending on the degree of membership, the plots do not report tick marks and labels on the y-axis. Family: S0 = single without children; S1+ = single with one or more children; U0 = in a union without children; U1 = in a union with one child; U2+ = in a union with two or more children; D0 = divorced without children; D1+ = divorced with one or more children. Employment: Educ = education; OLF = out of labour force; PT = part-time work; FT = full-time work.

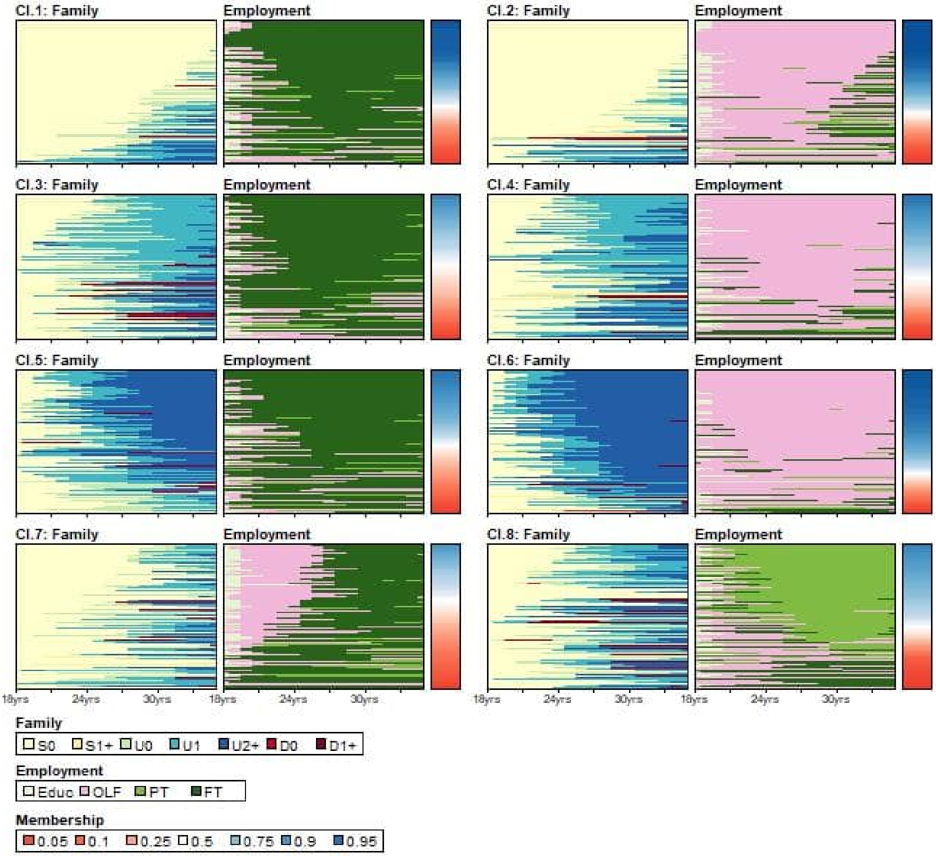
Figure A-3: Weighted gradient index plots for the six fuzzy clusters of the family and employment trajectories



Source: Multi-purpose Survey on Household and Social Subjects 2009.

Notes: For each cluster the sequences in the two domains are ordered on the vertical axes based on their degree of membership to the cluster; the height of the horizontal bar representing each sequence is proportional to the degree of membership. Since the sequences' bars differ depending on the degree of membership, the plots do not report tick marks and labels on the y-axis. For each cluster the gradient bar displays the degree of membership of each case to the cluster. Note that for each cluster all the sequences in the dataset are displayed: The differences across clusters are actually due to the different ordering of the sequences and to the different heights of the bars. Family: S0 = single without children; S1+=single with one or more children; U0 = in a union without children; U1 = in a union with one child; U2+ = in a union with two or more children; D0 = divorced without children; D1+ = divorced with one or more children. Employment: Educ = education; OLF = out of labour force; PT = part-time work; FT = full-time work.

Figure A-4: Weighted gradient index plots for the eight fuzzy clusters of the family and employment trajectories



Source: Multi-purpose Survey on Household and Social Subjects 2009.

Notes: For each cluster the sequences in the two domains are ordered on the vertical axes based on their degree of membership to the cluster; the height of the horizontal bar representing each sequence is proportional to the degree of membership. Since the sequences' bars differ depending on the degree of membership, the plots do not report tick marks and labels on the y-axis. For each cluster the gradient bar displays the degree of membership of each case to the cluster. Note that for each cluster all the sequences in the dataset are displayed: The differences across clusters are actually due to the different ordering of the sequences and to the different heights of the bars. Family: S0 = single without children; S1+ = single with one or more children; U0 = in a union without children; U1 = in a union with one child; U2+ = in a union with two or more children; D0 = divorced without children; D1+ = divorced with one or more children. Employment: Educ = education; OLF = out of labour force; PT = part-time work; FT = full-time work.