An Integrated Heuristic for Validation in Sequence Analysis

Raffaella Piccarreta

Department of Decision Sciences, "Carlo F. Dondena" Centre for Research on Social Dynamics,
Bocconi University, and BIDSa (Bocconi Institute for Data Science and Analytics)
Bocconi University
via Guglielmo Röntgen 1, 20136 Milano (Italy)
Telephone: +39 02 58365659
Email: raffaella.piccarreta@unibocconi.it

Emanuela Struffolino

WZB Berlin Social Science Center
Reichpietschufer 50
10785 - Berlin, Germany
Telephone: +49 30 25491 434
Email: emanuela.struffolino@wzb.eu

Abstract

Sequence Analysis is a collection of tools to describe life courses represented as sequences that are increasingly applied in different fields, particularly in demography, sociology, and political sciences. Identifying typologies through cluster analysis, thus disregarding individual sequences' peculiarities, is the aim of most applications. However, a substantive interpretation of such typology can be questionable when clusters include sequences deviating from the others. We propose an integrated approach to identify such sequences, distinguishing between sequences presenting structural peculiarities and randomly deviating sequences. We monitor the quality of partitions with respect to the amount and type of deviation in each cluster relying on novel graphical tools allowing to properly visualize and closely inspect the structure of deviating sequences. We demonstrate that the identification of deviating sequences provides relevant insights also when clusters are used as dependent or independent variables in an explanatory framework, for example in combination with multinomial logistic regression analysis.

Keywords: cluster analysis, validation, sequence analysis, visualization, index plot

An Integrated Heuristic for Validation in Sequence Analysis

Abstract

Sequence Analysis is a collection of tools to describe life courses represented as sequences that are increasingly applied in different fields, particularly in demography, sociology, and political sciences. Identifying typologies through cluster analysis, thus disregarding individual sequences' peculiarities, is the aim of most applications. However, a substantive interpretation of such typology can be questionable when clusters include sequences deviating from the others. We propose an integrated approach to identify such sequences, distinguishing between sequences presenting structural peculiarities and randomly deviating sequences. We monitor the quality of partitions with respect to the amount and type of deviation in each cluster relying on novel graphical tools allowing to properly visualize and closely inspect the structure of deviating sequences. We demonstrate that the identification of deviating sequences provides relevant insights also when clusters are used as dependent or independent variables in an explanatory framework, for example in combination with multinomial logistic regression analysis.

Keywords: cluster analysis, validation, sequence analysis, visualization, index plot

Introduction and Motivation

*If there is a figure connected with the problem [the student]
should draw a figure to point out on it the unknown and the data*

G. Polya

Eureka!

Archimedes

Sequence Analysis (SA) is a collection of tools increasingly used in social sciences to describe simultaneously the timing, the sequencing, and the *quantum* of processes represented by the sequences of events (states) experienced or occurred over a specific period, observed at regular intervals. Such sequences are regarded as meaningful units of analysis, and studied in their entirety. Stimulated by the recent widespread availability of longitudinal data, SA has been used extensively in demography and life course research. Typical applications concern, for example, family formation processes (Aassve, Billari and Piccarreta 2007; Struffolino, Studer, and Fasang 2016), intergenerational transmission of patterns of family formation (Fasang and Raab 2014), work-family trajectories (Aisenbrey and Fasang 2010), residential trajectories (Stovel and Bolan 2004), employment careers (Van Winkle and Fasang 2017); school-to-work transition (Brzinsky-Fay and Solga 2016; McVicar and Anyadike-Danes 2002). Even so, several applications in other fields (social policy analysis, democratization research, electoral participation studies, historical sociology, developmental psychology, mobility and time use research, and the study of cultural processes) prove the versatility of SA in analyzing the evolution of processes over time.

Rooted in the so-called algorithmic culture (Billari 2005; Breiman 2001), SA is not grounded on specific hypotheses about the generative process underlying sequences regularities (Abbott 1992). It therefore differs from other methods for

longitudinal categorical data (such as event history models, multi-state models, or latent Markov models) belonging to the statistical variable-centered or event-oriented culture, and relying on the assumption of a stochastic data-generative process. SA is, indeed, individual-centered: the sequences – regarded as realizations of a specific process – constitute the units of the analysis, and are studied as a whole in their temporal unfolding. SA aims at summarizing data to unveil their most common patterns, thus extracting information about the structural features of processes holistically understood, which have become almost invisible to the event-oriented methodological traditions (Abbott 1990, 1998).

In his seminal works Abbott (1990, 1995) focused on the (data-driven) individuation of ideal types of sequences. Specifically, he proposed to assess the dissimilarity between sequences by extending to social sciences the optimal matching (OM) algorithm, used in biology to analyze DNA strings. OM analysis calculates the dissimilarity between two sequences as the effort needed to transform one into the other. The resulting dissimilarities can be used as the input of a cluster analysis (CA) that partitions the data into groups representing alternative empirical realizations of a temporal process.

A lively season of criticisms towards pioneering works using OM in the social sciences (Levine 2000; Wu 2000), was followed by an equally lively season of methodological innovations and theoretical reasoning (Abbott 2016; Aisenbrey and Fasang 2010; Blanchard, Bühlmann, and Gauthier 2014; Piccarreta and Studer 2018; Ritschard and Studer 2018) aimed at addressing some of these concerns. Surely, one of the main issues that remain unaddressed so far is the validation of cluster solutions. Using CA to identify the most typical patterns in data remains in fact the goal – or at least a fundamental preliminary step – in the vast majority of the

applications of SA. Even when a taxonomy – that is a natural grouping of sequences – exists, the underlying structure that generated the data is typically unknown (see Warren et al. 2015 for discussion). This prevents validation in a statistical sense, because the ability of clusters to recover the “true” structure of data cannot be assessed. On one hand, the quality of a partition can be evaluated by referring to several criteria (not necessarily concordant), that can support decisions about the specific clustering algorithm to use or about the number of clusters to consider. On the other hand, even when a partition is satisfactory from the statistical point of view there is no guarantee that the identified types are the most relevant ones from a substantive point of view (with respect to the research question). It is therefore hardly possible to assess the sociological relevance of a specific solution, and the consistency across alternative options (Piccarreta and Studer 2018; Studer et al. 2011). This can be particularly relevant when clusters are used as dependent or independent variables in subsequent analyses. In many applications, data-driven types summarizing clusters’ features are *substituted* to the original (more turbulent) sequences, and multinomial logistic regression is applied to relate types to covariates (or, less frequently, to relate responses to types). This approach can be questionable and can lead to unreliable results when types do not properly represent all cases in the clusters. Even so, it is frequently adopted, mostly because of its simplicity and ease of interpretation, and because it allows considering the entire trajectories as response variables – thus (apparently) offering a clear interpretation of the effects of covariates on sequences.

We think that these flaws do not undermine the results obtained using SA in combination with CA, and should not discourage its use. A growing literature in different disciplines demonstrated the advantages of SA for the holistic study of

processes. However, we argue that attempting at identifying an “optimal” partition based on some specific criteria, even when well-grounded, is far from being an ideal strategy. In fact, the validity of any solution can always be questioned based on alternative criteria. We therefore suggest regarding a specific partition as a convenient “starting point” rather than as the ultimate goal of the analysis. We build on the idea that a critical inspection of the partition and of the clusters’ internal composition, aimed at identifying sequences that are inconsistent with their cluster and therefore might compromise its “validity”, allows a clear understanding of the data structure in its complexity, increasing the confidence on the conclusions that one can draw based on CA.

At this aim, we introduce criteria to identify *deviant* sequences in clusters, distinguishing between *structurally peculiar* sequences, having low-frequency features, and sequences with no clear or distinctive characteristics, which can be regarded as *random noise* (outliers). We rely on the amount of deviation within clusters both to compare partitions with a different number of clusters (thus pointing to the problem of properly selecting the number of clusters) and to identify possible rare types or sub-types in data, to get a more informed and clear description of the data structure. We introduce graphical tools to visualize the deviating sequences and to explore their features. We finally envision criteria to study the possible impact of the type and severity of deviations on the results obtained when clusters are used as dependent or independent variables in subsequent analyses, that is in an explanatory framework. We illustrate our proposals referring to data on employment trajectories of women in Italy.

Sequence Analysis and Cluster Analysis: What can we Learn from Deviant Sequences?

The application of SA requires some fundamental steps, each implying choices left to the researcher. First, each case is associated a sequence, that is the ordered collection of events (or states) experienced over a specific period. This requires selecting the states used to describe the life courses. A further crucial step concerns the choice of the dissimilarity criterion. OM analysis was initially proposed to quantify the consistency between two realizations of the hypothetical data's generative process by summarizing all the information available on the (portions) of the process represented by the sequences (Abbott and Forrest 1986). In OM, the dissimilarity between two sequences is based on the "cost" needed to align them using three operations: the substitution of one state with another, and the insertion or deletion (*indel* operations) of one state. To each operation specific costs are assigned, and the dissimilarity is the minimum total transformation cost, computed using the Levenshtein matrix for alignment (Levenshtein 1966; Macindoe and Abbott 2004). The original proposal has been largely discussed (see Studer and Ritschard 2018 for review). Initially, the attention was pointed to the choice of the costs assigned to each operation. Subsequently, alternative criteria were introduced that differ with respect to the possible aspects (sequencing, timing, or duration) that one could emphasize when assessing the extent of dissimilarity between trajectories.

The dissimilarities – obtained based on a given criterion – are then used to apply data-reduction techniques, typically CA. Among the several clustering algorithms available based on dissimilarities, the most famous and used ones are hierarchical algorithms – that sequentially join cases – and partitioning algorithms – that iteratively assign cases to clusters so as to (locally) maximize the clusters' within

homogeneity. A further crucial issue concerns the choice of the number of clusters. Several indices are available to monitor the quality of partitions with a varying number of clusters (e.g., Point Biserial Correlation, Average Silhouette Width, Pseudo R^2 and Pseudo F statistic, to cite some of the most used; see Hennig and Liao 2010 and Studer 2013 for detailed review and discussion). Eventually, different choices typically lead to different solutions, even if the ultimate goal of the entire procedure is the same, namely to identify the most typical patterns in data disregarding the possible individual sequences' peculiarities, and thus considering the difference between types and individual realizations as a random non-informative error term (Studer et al. 2011).

While acknowledging the relevance of each step, here we focus specifically on two issues related to the focal problem of identifying typical patterns using CA. The first concerns the identification of the number of types in data. Actually, every data-driven partition might fail to uncover the “true” number of types. On the one hand, it might be a simplification combining different types into the same cluster or assigning cases of a given type to other types. On the other hand, the partition might be unnecessarily complicated because cases belonging to the same type are separated into smaller, distinct clusters. In addition, since empirical analyses rely upon samples, it might happen that some types are not frequent enough to be identified and isolated in dedicated clusters, unless a relatively large number of clusters is extracted.

A second problem relates to the assignment of cases to the cluster they “truly belong”. Even when the number of types is correctly identified, some sequences can deviate consistently from their cluster. In some cases, such sequences belong to sub-types, or to types with relatively low frequency. In other cases, instead, they are

outliers, i.e. sequences presenting very peculiar and uncommon features. In both situations, clustering algorithms will forcedly assign these sequences to clusters, although they have little in common to the other cases, and the ideal-types signified by the clusters do not account for the possible differences within clusters.

Unfortunately, the true structure underlying the data is unknown, and therefore it is not possible to validate one specific solution by assessing if and to what extent it recovers the data structure. We propose an integrated heuristic approach for cluster validation, allowing an in depth analysis aimed at identifying the possible types of sub-patterns within clusters, separating “regular” from deviating cases. “Regular” cases in clusters are those reflecting the most relevant types. Deviating cases can be either structurally different from most cases in the cluster or they have very specific and uncommon patterns, which can therefore be considered as outliers. We refer to the first type of deviant sequences as *structurally peculiar*, and to the second type as *random noise*.

The identification of the different types of deviating trajectories can be useful under different respects. First, it permits a more effective exploration of the characteristics both of the “regular” and of the deviant sequences, and allows a better description of the types, preventing the possible influence of deviating sequences on them.

Second, it offers insights on the “quality” of the within-clusters heterogeneity, i.e., information on the type of deviations from the main patterns identified by the clusters.¹ If many sequences in a cluster have peculiar distinctive features, this might indicate that the chosen partition failed in identifying the “true” number of groups in data. Under this perspective, the study of the presence and of the severity of deviations offers a valuable additional criterion to monitor the quality of partitions with

a varying number of clusters. This can hopefully support the researcher in selecting an appropriate cluster solution, and inform a more conscious and cautious interpretation of the substantive results.

Finally, an in-depth analysis of the clusters' composition might alleviate the relevance of the choices made by the researcher with respect to the dissimilarity criterion and to the clustering algorithm, because, as mentioned before, a specific clustering solution is regarded as a convenient "starting point" rather than as the ultimate goal of the analysis. Indeed, in the procedure streamlined in the next section, we use CA as a genuinely explorative tool to better grasp the structure of the data before even planning any following analyses. The identification of deviant sequences can prove particularly useful when clusters enter as dependent or independent variables in subsequent analyses. Evaluating the extent to which the obtained results depend on the presence and type of deviating sequences enhances the rigorousness of these analyses, and supports the substantive conclusion based on them.

Rather than on statistical tools or criterion, our approach for clusters' validation attempts at identifying all the *relevant* groups in data, and relies on a heuristic that leverages on two of the pillars of SA, namely the measurement of the dissimilarity between sequences and the effective exploration of trajectories through suitable visualization tools.

Identifying Deviating Sequences: Our Proposals

We regard as *deviating* those sequences that differ from the others in their cluster with respect to the most relevant characteristics of the process signified by the cluster itself or, in other words, by the *type* summarizing the features of sequences

allocated to the cluster. Deviation can arise because sequences are structurally peculiar, and the chosen partition is not detailed enough to include an additional type they would be better allocated to. Otherwise, sequences deviate because they have very uncommon traits that cannot be assimilated to those of other sequences. In what follows, we introduce some criteria to identify both types of deviation. In Appendix B, we briefly review some possible alternative strategies that might apparently serve the purpose of identifying deviating sequences, and discuss the reasons why we decided not to consider them.

Dissimilarity of Sequences from their Cluster

A first criterion to identify deviating cases refers to the relative extent of their dissimilarity from their cluster. This can be evaluated considering the average dissimilarity of each sequence from all the others in its cluster or its dissimilarity from the cluster's representative sequence, which can conveniently be the medoid, that is the case in the cluster with the minimal average distance from all the other cases. Thus, for a sequence s_i placed in a cluster C_g including n_g cases summarized by a medoid $\tilde{s}_g$ we consider:

$$a_i = \sum_{j \in C_g, j \neq i} d(s_i, s_j) / (n_g - 1)$$

$$m_i = d(s_i, \tilde{s}_g).$$

Individual deviations can be compared with general or clusters-specific summaries. One possibility is to use as references the averages calculated across all the cases, that is $\bar{a} = \sum_i (a_i / n)$ and $\bar{m} = \sum_i (m_i / n)$, thus avoiding a definition of deviation excessively related to the clusters' specific characteristics. It is also possible to refer to deviations standardized using the entire sample's or the clusters' standard

deviations. In any case, it is necessary to define suitable thresholds, which nonetheless might depend on the level of noise in data. Therefore, we postpone a discussion about the operational aspects related to the calculation of the two indices until the illustration of the results obtained by applying these criteria to our data.

Identification of Core, Border and Noise Sequences

The two measures described so far are expected to individuate sequences deviating from their cluster because of specific characteristics distinguishing them from the others, such as for example prevalence of rare states, or length of states duration. Such sequences might reflect some sub-types or rare types in data, and they are not necessarily isolated. Indeed – at least in principle – one could expect to see at least some of such sequences placed in dedicated clusters when the number of clusters is high enough.

It is nonetheless important to envision criteria pointing to the individuation of outlying sequences, isolated from all the others. Thus, while the previous measures allow exploring the “tails” of the distributions of the sequences within clusters, we borrow from the DBSCAN clustering algorithm (Ester et al. 1996) to individuate outliers. DBSCAN classifies cases into core, border and noise, depending on whether they lay in “dense” regions, at their borders, or “far” from them. More precisely, a case is classified as “core” if the number of its neighbors – i.e. cases lying within a pre-defined distance from it – is higher than a specific threshold. Instead, a border case is a non-core case lying in the neighborhood of a core case; a noise case is one that is neither a core nor a border point. This approach relies on the specification of the radius of the neighborhood and of the minimum number of cases it should include. At this aim, we consider for each sequence the dissimilarity

from its 5th nearest neighbor (NN) in its cluster, d_i^{NN5} . We refer to the distribution of the d_i^{NN5} 's in the entire sample, and focus on one of the highest percentiles, say the 95th, $\bar{d}_{0.95}^{NN5}$. Thus, the 95% of the sequences in the sample have at least 5 NN (in their cluster) within a distance lower than or equal to $\bar{d}_{0.95}^{NN5}$. We define as core sequences those with at least 5 sequences in their cluster within a distance $\bar{d}_{0.95}^{NN5}$, and border and noise sequences accordingly. Substantially, noise sequences are relatively isolated, and are therefore potential outliers. Observe that it is necessary to choose the number of NN and the percentile of the distances to be used as threshold. Nonetheless, conditioned to these choices, the procedure is entirely automatic.

Qualifying and Exploring Within-Clusters Heterogeneity: Heuristics Based on Deviant Sequences

Deviating Sequences and Number of Clusters

The amount, the types, and the severity of deviations in clusters can inform the final decision of the researcher on the number of clusters. Indeed, the presence of structurally different sequences can suggest the opportunity of extracting additional clusters; also, the amount of noise in clusters can be used to monitor the quality of partitions of different degree.

Besides the criteria commonly used to choose the number of clusters, we propose to analyze the structure of the sequences responsible of the observed within-cluster dispersion. If alternative cluster solutions lead to the identification of similar cores and mostly differ with respect to where deviating sequences are placed, this might enhance the confidence in the existence of a structure in the data. In other words, from a substantive point of view this would permit to evaluate whether some specific sub-types are identified and aggregated to the same clusters, and whether

their characteristics can be reasonably assimilated to those of the regular sequences.

Visualization of Deviating Sequences

To explore the features of deviating sequences within each cluster, we build on the *sequence index plot* (Scherer 2001), which is typically used to visualize sequences (possibly separately for each cluster). In this plot, cases are arranged on the vertical axis, and each sequence is represented by a horizontal line, which changes color depending on the visited state (alternatively, sequences can be arranged on the horizontal axis and represented by vertical lines). Typically, cases are sorted and arranged in the plot according to the time when a specific state was experienced for the first time, thus disregarding all the other – possibly relevant – differences among sequences. To improve the visualization, Piccarreta and Lior (2010) suggest to order the sequences according to their score on the first factor derived by applying Multidimensional Scaling (MDS) to the dissimilarity matrix. Even so, the plot might fail to accurately describe the patterns in data when the sample size is large, because of the tendency to overplot the sequences to make them fit into the graphical window. In this case, visualizing and distinguishing regular and deviant sequences within each cluster and identifying their features become almost infeasible.

Under this perspective, the alternative plots introduced in the literature to avoid – or at least to limit – overplotting (see e.g., Fasang and Liao 2014; Gabadinho et al. 2011; Müller et al. 2008; Piccarreta 2012) are not useful, because they display only sequences that are representative in some respect, and therefore surely not deviating. Instead, we want to uncover the features of the most frequent patterns

without losing information on the least relevant – or deviating – ones. We therefore introduce the *flagged index plots*, which separately display regular and deviating sequences. Each cluster is associated a set of index plots, one dedicated to the regular cases and other/s dedicated to sequences deviating according to one (or more) definition or according to the severity of deviation, this allowing a close inspection of not regular sequences.

To further enhance the visualization of the sequences' features in our flagged index plots, we order cases resorting to seriation algorithms (Hahsler and Hornik 2011; Hahsler et al. 2008). In particular, we obtained satisfactory results using the so-called *Traveling salesperson problem* (TSP; Gutin and Punnen 2007). This optimization algorithm aims at finding the shortest path (in terms of dissimilarities) connecting all the cases (starting from one randomly chosen case).

Deviating Sequences and Multinomial Logistic Regression on Clusters

What discussed so far becomes even more relevant when the original sequences are substituted with the types describing their clusters, and multinomial logistic regression is used to relate the probability of being in each cluster to a set of covariates. This procedure is suitable only provided the clusters are sufficiently homogeneous, so that the simplification implied by the substitution of the original sequences with their cluster membership indicator is reasonable and appropriate. Taking another perspective, the possibly deviating sequences in the clusters should not influence (too much) the results obtained using multinomial logistic regression. Thus, even when one considers the assignment of deviating sequences to the clusters as reasonable or meaningful, it is important to assess whether deviating sequences are similar to their cluster with respect to the covariates. At this aim, we

suggest analyzing the sensitivity of the regression's results to different types and to different degrees of severity of deviations. This would undoubtedly strengthen the considerations and the conclusions based on these analyses.

An Application to the Employment Trajectories of Women in Italy

To illustrate our proposals we use data from the “Multi-purpose Survey on Household and Social Subjects” carried out in 2009 by the Italian National Statistical Office (ISTAT). The survey collects a wide range of 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.

As the work trajectories of men and women present structural differences – think for example to the higher propensity of women to choose part-time work or to exit the labor market when they have children – we limit attention to women. In particular, we focus on women born between 1959 and 1974, whose work activities could be tracked between the ages 16 and 35. Since the information on the month of birth is not available in the data, we left-aligned the trajectories at the first month in education, namely when respondents started elementary school. Specifically, we selected women who started elementary school between ages 5 and 7 (189 women who started school at an older age were excluded) in the months of September, October, or November (164 women were excluded).² This adjustment has the advantage to align individuals according to the moment when they were able to make the same decisions (e.g., continuation of education, entrance in the labor market), irrespective of the moment when they started school.

To define the individual sequences we consider the monthly labor activities (states) experienced from 120 to 350 months after the start of the school (thus from about the age of 16 to the age of 35 for the great majority of women, who started school at the age of 6). We consider a quite detailed set of states, distinguishing between education (Educ), Joblessness (including both unemployment and inactivity), Full-time (FT), and Part-time (PT) work. We further distinguish the working activities according to the type of contract: Permanent (Perm), Temporary (Temp), Self-employment (Self), or Dependent Self-employment (DepSelf).³

Our primary goal is to identify the most typical employment trajectories, and to evaluate their relationship with some (baseline) covariates, measured when the respondent was 15 years old. Specifically, we consider birth cohort (1959-1964, 1965-1969, or 1970-1974), macro-area of residence (North-West, North-East, Center, South, or Islands), highest parental education (primary, lower secondary, upper secondary, or tertiary), and working status of the mother (employed or not).⁴ The final sample consists of 4,323 women with complete information on all the relevant variables (please refer to Appendix A, Table A1, for the distribution of the covariates in the final sample).

Preliminary Choices in Sequence and Cluster Analysis

While acknowledging the importance of the choice of the dissimilarity criterion, we underline that such a choice is strongly dependent on the research question (see Piccarreta and Studer, 2018), and we therefore refrain from a comparison of results obtained using different criteria. To illustrate our proposals, we refer to the OM algorithm. Despite some criticisms, OM is without any doubt one of the most popular and used approaches, and it typically leads to dissimilarities nicely highlighting the

most relevant differences in trajectories with respect to experienced states, their durations and times at transitions from one state to another. Following a rather standard approach, we set the substitution costs between two states inversely proportional to the transition frequencies from one state to the other, and the insertion-deletion costs equal to 1.

As for the clustering algorithm, we point out that hierarchical algorithms are characterized by a certain rigidity, which does not allow revising the partition in the case of ‘bad choices’ in the first agglomeration steps. This can be unsatisfactory with noisy data, and we therefore prefer partitioning algorithms. Specifically, we use the Partitioning Around Medoids (PAM; Kaufman and Rousseeuw, 2005) algorithm, an extension of the *k*-means algorithm to the case when only dissimilarities are available. PAM attempts at iteratively assigning cases to a pre-fixed number of clusters, by minimizing the distances between cases in a cluster and the cluster’s medoid.

Several criteria are available to monitor the quality of partitions with a varying number of clusters. For the sake of completeness, here we use a battery of criteria: Point Biserial Correlation, Hubert’s Gamma, Hubert’s C coefficient, Average Silhouette Width, Pseudo R^2 and Pseudo F statistic – the latter calculated based both on the original and on the squared dissimilarities (see Hennig and Liao, 2010; Studer, 2013). The joint analysis of such indices can point to a reasonable range of solutions, that can be further analyzed and compared under different points of view, focusing for example on the substantive relevance, the specific interest of the obtained groups, or on the amount of heterogeneity within clusters.

Figure 1 reports the plots of the considered criteria for PAM-partitions with different number of clusters. Almost all the indicators present a local optimum

corresponding to 4 or 6 clusters, with a slightly higher support for the 4-clusters solution. This is the solution one would select based on the Average Silhouette Width (ASW) coefficient, one of the most popular in the applications on SA.

Please insert Figure 1 here

Comparing Partitions with a Different Number of Clusters: Insights from Deviating Sequences

The identification of deviating sequences within clusters can inform choices about a suitable (or convenient) number of clusters. For each case, we calculate the distance of cases from their cluster using the average distance from the cluster or the distance from the cluster's medoid. In addition, we identify noise sequences, referring to $\bar{d}_{0.95}^{NN5}$, the 95-th percentile of the distribution of the radius of the set of 5-NN calculated for all the cases in the sample. Figure 2 illustrates the relevance of deviating sequences within each cluster for partitions with a number of clusters between 2 and 10. In particular, for each cluster we monitor the average dissimilarity of cases from their cluster's medoid or from the other cases in their cluster, the percentage of cases with a standardized dissimilarity higher than 1.5 from their medoid and from the other cases, and the percentage of core and noise points. Notably, the 4-clusters solution, which turned out to be one of the most suitable according to the criteria reported in Figure 1, includes one cluster with a relatively high number of deviating sequences. As the number of clusters increases, we observe an expected relatively lower level of within-group heterogeneity. Moreover, the dissimilarities from sequences to their cluster decrease at a lower rate starting from the 6-clusters solution, with a stabilization of the most critical clusters'

heterogeneity when 9 clusters are extracted. Similar considerations arise when focusing on the percentages of core and noise points. On the one hand, these results seem to offer further support to the 6-clusters solution over the simpler partition with 4 clusters. On the other hand, they seem to suggest that the 9-clusters partition is a good compromise between complexity and noise.

Please insert Figure 2 here

It is important to underline that our criteria are *partition*-specific. Indeed, the standardized dissimilarities and the core/noise points were obtained based on summaries calculated on the deviations observed for the actual clusters in each partition. As the number of clusters increases, the standard within-clusters heterogeneity will generally decrease, and – for example – cases that were flagged as core points in a lower-degree partition could be flagged as noise once more compact clusters are considered. Nonetheless, our goal here is not to offer a new procedure to “automatically” select the number of clusters but, rather, to introduce a heuristic that provides researchers with additional tools for an in-depth exploration of the data structure.

At this aim, we also consider some plots that allow qualifying the within-clusters heterogeneity with respect to the types of deviations in clusters, and that can be used to compare the different partitions from a substantive point of view, i.e. with respect to the extracted types. Based on the evidence provided by Figures 1 and 2, we focus on the solutions with a number of clusters ranging between 4 and 9. The top panel in Figure 3 reports the index plot of the entire set of sequences. Since the number of cases is much higher than the number of time points used to build the

sequences, we place individuals on the horizontal axis. The plot therefore displays a sequence of vertical bars, each representing an employment trajectory. In the bottom panel, each sequence is associated a vertical bar with colors changing according to the cluster membership for the considered partitions. To enhance the visualization of the sequences and of the clusters' compositions, we applied a hierarchical algorithm to the medoids of the 9-clusters. We used information on the agglomeration history to order the 9 clusters along the horizontal axis, subsequently ordering sequences within each cluster using the TSP seriation algorithm.

Please insert Figure 3 here

Results in Figure 3 suggest that the “core” cases in the 4-clusters partition describe careers dominated respectively by (long periods in) full-time permanent jobs, unemployment, part-time permanent jobs, and self-employment. Such clusters nonetheless include some trajectories with less frequent or anomalous distinguished features. Moving to 5 clusters, a new cluster is obtained mostly including women originally the first cluster (the darkest shade of brown) who achieved full-time permanent contracts relatively later. Interestingly, moving to 6 and then to 9 clusters some trajectories (those on the right side of the plot on top) are grouped into dedicated clusters. Thus, refining the partition allows identifying types dominated by non-standard employment, namely part-time and temporary job, which are likely to be of substantive interest when looking at women’s labor market participation across cohorts. . The plot also shows that other trajectories – for example, those presenting relatively long periods with FT-Depself contracts – remain scattered across the clusters despite refinement. Nonetheless, it is not easy to individuate in the plot

(even if optimized using the TSP seriation algorithm) sequences that are isolated or have very specific features (being for example more turbulent than others, not dominated by a specific state, or presenting rare transitions or times at transitions), and this holds also when clusters-specific index plots (not reported) are used.

Visualizing Deviating Sequences

For a better identification of the features of deviating sequences, we explore clusters' compositions using the *flagged index plots*. We start analyzing the 4-clusters partition. Within each cluster, we flag as deviant cases with an average dissimilarity from the others (in their cluster) or from their clusters' medoid higher than 1.5 times the (respective) general average, and cases identified as noise by the $\bar{d}_{0.95}^{NN5}$ -criterion. In Figure 4, clusters are displayed row-wise, and each cluster is described by one plot (on the left) including only cases that are flagged as regular by all considered criteria, and by three plots reporting the sequences flagged as deviant by each of the considered criteria (note that these sets of sequences might overlap). The most relevant patterns in each cluster – evident in the plots of regular cases – coincide with those deduced from Figure 3. Nonetheless, the isolation of deviating sequences in dedicated plots makes it evident that cases with long time-spans spent in FT-Temp, FT-Depself, and PT-Temp do not belong to the types emerging from the four clusters. Also, the sets of cases dissimilar from their cluster (Figure 4, columns a and b) mostly include *blocks* of sequences sharing some distinguishable patterns. These “structurally peculiar” sequences identify sub-types or less relevant types, experienced only by few women. This is likely to reflect the relatively small diffusion of the employment arrangements dominating these trajectories (particularly FT-Depself and PT-Depself, as well as FT-Temp and PT-Temp) among Italian women

(at least during the considered age period). In fact, one could reasonably expect to see these cases placed in dedicated clusters when a sufficiently high number of clusters is extracted. Instead, the set of sequences flagged as noise (Figure 4, column c), includes both turbulent sequences, with very uncommon characteristics (e.g., high number of transitions or rare transitions), and sequences with more regular – even if atypical – patterns, dominated by extremely rare states (particularly, PT-Depsel). These cases cannot be described by a common or relevant (in terms of frequency) pattern, and therefore can be regarded as outliers or – more in general – as irregular or unstructured deviations from the types signified by the extracted clusters.

Please insert Figure 4 here

These considerations hint at alternative cluster solutions, for example the 6- and the 9-clusters partitions suggested by the monitoring criteria in Figures 1 and 2 and by results in Figure 3. Figure 5 reports the flagged index plots for the 9-clusters partition (built following the same criteria to flag deviant cases). The comparison of the regular cases in the 4-clusters and in the 9-clusters partitions strengthens the evidence offered by Figure 3, and better qualifies the refinement provided by the latter solution. For example, sequences dominated by FT-Temp are now grouped in a dedicated cluster (the fourth), while sequences characterized by other less frequent states are still flagged as deviant.

Please insert Figure 5 here

This indicates that even if the number of clusters is increased from 4 to 9 the less frequent types would not be isolated and identified. Alternative flagged index plots could be employed to compare different levels of severity of deviations, using different thresholds for the proposed criteria (dissimilarity from cluster and percentile of the distribution of the 5-NN radius). For our application, using different thresholds (see Fig. A2, A3, and A4 in Appendix A) did not offer additional insights about the structure of the data.⁵

The flagged index plots allow distinguishing between irregular and structural deviations, to highlight their features, and to analyze how – or based on what characteristics – they are aggregated to the clusters. In fact, a certain degree of deviation from the “regular” type signified by the clusters is expected, due to several factors affecting individuals’ behaviors (e.g. preferences, resources, constraints). Nonetheless, *structurally* different sequences cannot be interpreted or regarded in light of this consideration, because their aggregation to clusters is mostly related to the chosen degree of the partition, and should be justified based on some substantive (in this case, sociological) considerations. This is not always possible. For our data, for example, structurally deviant sequences with similar characteristics are placed in different clusters, mainly based on the length of education.

It is worth to point out once more that CA only leads to a reasonable reorganization of the information in data. A final choice among competing solutions should be necessarily based on sociological evaluations, and specifically on how reasonable or interpretable are the identified types and/or how relevant are the sub-types in clusters when the number of groups is increased.

In several applications, the problem related to the presence of sequences dominated by states with a relatively low frequency is addressed by simplifying the

states used to build the sequences. For example, for our data, different simplifications are possible: one could avoid distinguishing between full- and part-time work, or could simplify the type of contract. Nonetheless, this should be grounded on substantive reasons, because the induced decrease in data heterogeneity will typically lead to the identification of a stronger cluster structure, with less deviant sequences (and probably with higher ASW). However, this procedure can be questionable when *structurally* deviant sequences are detected. As we showed when comparing the 4- and 9-cluster solutions, such sequences can be grouped in a new cluster, and lead to a new type, provided that a sufficiently high number of clusters is extracted. We argue that the heuristic proposed here can support better informed choices in this respect. In fact, it seems advisable to build sequences using a detailed set of the states. An accurate analysis of the features of deviant sequences would substantiate and support or rather discourage the decision of recoding less frequent states assuming that they are equivalent to others.

A different strategy consists in removing deviant sequences from the set of data. This can be reasonable when one wants to highlight the most relevant features of the cluster and/or wants to extract clusters' representative, for example the medoids, preventing the influence of deviant sequences on them. The removed sequences can be grouped in a catchall category that cannot be substantively interpreted (Lesnard, 2014), or further clustered to unveil substantively meaningful sub-types that did not emerge from the initial partitioning. In this respect, our heuristic allows the researcher to make explicit the criteria used to identify regular cases and to separate them from the others.

Deviating Sequences and Multinomial Logistic Regression: Sensitivity Analysis

Further evaluations are necessary when employing procedures that go beyond the identification of types, and use cluster membership as an input in subsequent analyses. In fact, in many applications multinomial logistic regression is applied to test whether cluster membership is related to baseline covariates. In this situation, it seems important to evaluate if structurally deviant sequences have also structurally different characteristics with respect to the explanatory variables. In fact, when a strong relationship exists between sequences and covariates one can reasonably expect structurally peculiar trajectories with distinct behaviors with respect to the covariates' levels, with a consequent possible impact on regression. The aggregation of such structurally deviant sequences to clusters dominated by other types might indeed mask the relevance of specific regressors.

As for the noise sequences, they are typically very different one from another. Thus, one could reasonably expect that their deviation from the other cases will be unstructured and “random” also with respect to covariates. Nonetheless, some of these sequences might present outlying combinations of the covariates' levels, and their possible impact on the regression's results will possibly depend on the severity of their deviations. Arguably, this is relatively unusual when few covariates – possibly binary or with few categories as in our and most applications – are taken into account. Therefore, even if in some situations noise sequences can bias the results of regressions, the probability of incurring in such a risk is typically higher when structurally deviating sequences are present. In any case, an analysis of the sensitivity of regression results to the type and severity of deviation in data seems necessary to guarantee the reliability of results.

For the data at hand, we apply a multinomial regression model predicting cluster membership based on a set of covariates measured at age of 15, that is before the start of the observational window on employment trajectories. For illustrative purposes, we perform this analysis on the 4-cluster solution, specifically because it would be the partition that one would select based on the commonly adopted ASW coefficient.

Figure 6 reports the confidence intervals of the coefficients of a battery of multinomial logistic regressions predicting the probability of being assigned to each cluster (reference category: cluster 2, dominated by unemployment) focusing on different samples. The first includes all the cases (in black), the others exclude respectively cases dissimilar from their cluster (identified either as average-dissimilar from their cluster or dissimilar from their cluster's medoid, in red), noisy cases (in blue), and both the types of deviating cases (results in green). Note that in some cases the coefficients are indeed affected by deletion of deviating cases. Interestingly, as expected, this is especially true when sequences dissimilar from their cluster are filtered out. In some cases, the results for the covariates lose significance (consider for example the binary variable indicating the mother's working status, M.Occ) or hint at a possible change in the direction of the relationship between clusters and covariates (as for the coefficients for the dummy variable indicating a tertiary highest parental education level, PEdu4).

Please insert Figure 6 here

Our results suggest that limiting attention to the standard quality criteria presented in Figure 1 might be misleading when the actual interest is to identify a

stable structure in the data with respect to their association with the covariates, or – in other words – when we want to assess the homogeneity of the clusters with respect to the covariates. Note that this sensitivity analysis can also be performed flagging deviant cases using different thresholds, in an attempt to evaluate the impact of an increasing severity of deviation on the regression coefficients.

Conclusions

“The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage” (Jain and Dubes 1988: 222). The integrated heuristic illustrated in this paper represents a step forward in the direction of validation of CA when used in SA applications aimed at identifying typologies, whose substantive interpretation can be questionable when clusters include sequences deviating from the others. Our strategy for the individuation of different types of deviating sequences combined with graphical tools for their visualization provide researchers with a toolkit rich enough to explore the data structure, and offers additional insights to select a partition adequately describing it. In addition, it permits to evaluate the reliability of the results obtained using multinomial logistic regressions based on clusters.

Our approach for validation relies on the criteria used to flag deviant sequences, as well as on the chosen thresholds. For example, in our application we compared the dissimilarity of sequences from their clusters with the general average, but other summaries could be considered, for example the percentiles. Also, one could refer to clusters-specific summaries or thresholds, depending on clusters’ turbulence or complexity. We underline that our goal is not to set rules, because there is not a

unique way to explore deviations, and different approaches can be adopted depending on the characteristics of the data at hand.

Our application suggests that even when a relatively large number of clusters is extracted, there might be sequences poorly related to their cluster. Under this perspective, we share with Warren and colleagues (2015) the concerns about the risk associated with over-reliance on the results obtained using CA. Even so, we address these risks through a proper exploration of the data in its complexity, and a proper consideration of the possible excess of simplification of a given partition. This is surely important both when the goal is unveil typical patterns and when all the sequences in the same cluster are treated as realizations of the same “type”, and cluster membership is used as input in subsequent analyses.

Our proposals can also be employed to explore the possible presence of stable structures in data by comparing results obtained using different approaches to cluster cases. In fact, one could reasonably expect to observe similar regular cases in clusters extracted using alternative clustering method, especially when the number of clusters increases. The same holds for clusters obtained based on alternative dissimilarity criteria, unless a measure focused on very specific features of the sequences is used.

On the one hand, the aim of the analysis and the sociological a-priori assumptions or hypotheses will – and should – guide the researchers in the choice of the final partition and of the approach to manage cases deviating from the identified types. On the other hand, using CA to unveil the possible structure in data weakens the reliance of results on technical choices, e.g. clustering algorithm or criteria to select the number of clusters, and allow researchers to make informed decisions about the final partition to take into account. A clear explanation and

discussion of the approach followed to deal with data complexity would prove useful in clarifying how empirical evidence is used to identify types, and would definitely increase the confidence on the conclusions drawn based on CA.

Notes

1. Note that this holds also for soft clustering algorithms. For example fuzzy clustering algorithms assign a vector of weights to each sequence indicating the strenght of its membership to each cluster. Also in this case, some sequences could present almost constant weights (being therefore unrelated to the obtained clusters) or present a relatively higher closeness to one specific cluster without a clear or strong sharing of its main features.
2. Although in most recent decades the school in Italy starts in September in all the regions, this was not the case when the older cohorts of women entered the educational system (namely, the 1960s and the 1970s).
3. In Italy, the deregulation reforms of the early 1990s introduced para-subordinate collaboration. Employees with this type of contracts are also named dependent self-employed. In fact, they are self-employed in legal terms, since they are not formally subordinated to an employer. Nonetheless, they are often 'economically dependent' on an employer because, in most cases, their activity relies upon the clients provided by the employer.
4. We do not consider the occupation of the father because it is almost constant across cases.
5. A variation of the flagged index plot allows isolating from the intial clusters of a chosen partition the sequences removed from each type when a higher number of

clusters is extracted. This visualization tool is suitable only for hierarchical partitioning: Figure A1 in the Appendix displays an example of this split-plot.

References

- Aassve, Arnstein, Francesco C. Billari, and Raffaella Piccarreta. 2007. "Strings of Adulthood: A Sequence Analysis of Young British Women's Work-Family Trajectories." *European Journal of Population/Revue Européenne de Démographie* 23(3-4):369-388.
- Abbott, Andrew. 1990. "A Primer on Sequence Methods." *Organization Science* 1(4):375-392.
- Abbott, Andrew. 1995. "Sequence Analysis: New Methods for Old Ideas." *Annual Review of Sociology* 21(1):93-113.
- Abbott, Andrew. 1998. "The Causal Devolution." *Sociological Methods & Research* 27(2):148-181.
- Abbott, Andrew. 1992. "From Causes to Events Notes on Narrative Positivism." *Sociological Methods & Research* 20(4):428-455.
- Abbott, Andrew. 2016. *Processual Sociology*. Chicago: Chicago University Press.
- Abbott, Andrew and John Forrest. 1986. "Optimal Matching Methods for Historical Sequences." *The Journal of Interdisciplinary History* 16(3):471-494.
- Aisenbrey, Silke and Anette Eva Fasang. 2010. "New Life for Old Ideas: The 'Second Wave' of Sequence Analysis Bringing the 'Course' Back Into the Life Course." *Sociological Methods & Research* 38(3):420-462.
- Billari, Francesco C. 2005. "Life Course Analysis: Two (Complementary) Cultures? Some Reflections With Examples From the Analysis of the Transition to Adulthood." Pp. 267-88 in *Owds an Interdisciplinary Perspective on the Life*

Course, edited by Levy, René, Paul Ghisletta, Jean-Marie Le Goff, Dario Spini, and Eric Widmer.

Blanchard, Philippe, Felix Bühlmann, and Jacques-Antoine Gauthier (editors). 2014. *Advances in Sequence Analysis: Theory, Method, Applications*. Vol. 2. Cham Heidelberg New York Dordrecht London: Springer International Publishing.

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16(3):199-231.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees*. Wadsworth Inc.

Brzinsky-Fay, Christian and Heike Solga. 2016. "Compressed, Postponed, or Disadvantaged? School-to-Work-Transition Patterns and Early Occupational Attainment in West Germany." *Research in Social Stratification and Mobility*.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." Pp. 226-231 in *Kdd*. Vol. 96.

Fasang, Anette Eva and Marcel Raab. 2014. "Beyond Transmission: Intergenerational Patterns of Family Formation among Middle-Class American Families." *Demography* 51(5):1703-1728.

Gutin, Gregory and Abraham. P. Punnen (editors). 2007. *The Traveling Salesman Problem and Its Variations*. Springer US.

Hahsler, Michael and Kurt Hornik. 2011. "Dissimilarity Plots: A Visual Exploration Tool for Partitional Clustering." *Journal of Computational and Graphical Statistics* 20(2):335-354.

- Hahsler, Michael, Kurt Hornik, and Christian Buchta. 2008. "Getting Things in Order: An Introduction to the R Package Seriation." *Journal of Statistical Software* 25(3):1-34.
- Hennig, Christian and Tim F. Liao. 2010. "Comparing Latent Class and Dissimilarity Based Clustering for Mixed Type Variables with Application to Social Stratification, Technical Report."
- Kaufman, Leonard and Peter Rousseeuw. 2005. *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken: John Wiley & Sons.
- Levenshtein, Vladimir I. 1966. "Binary Codes Capable of Correcting Deletions, Insertions and Reversals." *Soviet Physics Doklady* 10:707-710.
- Levine, Joel H. 2000. "But What Have You Done for Us Lately? Commentary on Abbott and Tsay: Sequence Analysis." *Sociological Methods & Research* 29(1):34-40.
- Macindoe, Heather and Andrew Abbott. 2004. "Sequence Analysis and Optimal Matching Techniques for Social Science Data." Pp. 387-406 in *Handbook of Data Analysis*, edited by Melissa Hardy and Alan Bryman. Thousand Oaks, CA: Sage.
- McVicar, Duncan and Michael Anyadike-Danes. 2002. "Predicting Successful and Unsuccessful Transitions from School to Work by Using Sequence Methods." *Journal of Royal Statistical Society* 165(2):317-334.
- Piccarreta, Raffaella and Orna Lior. 2010. "Exploring Sequences: A Graphical Tool Based on Multi-Dimensional Scaling." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 173(1):165-184.
- Piccarreta, Raffaella and Matthias Studer. 2018. "Holistic Analysis of the Life Course: Methodological Challenges and New Perspectives." *Advances in Life Course Research*.

- Ritschard, Gilbert and Matthias Studer (editors). 2018. "Sequence Analysis and Related Approaches Innovative Methods and Applications." Cham: Springer.
- Scherer, Stefanie. 2001. "Early Career Patterns: A Comparison of Great Britain and West Germany." *European Sociological Review* 17(2):119-144.
- Stovel, Katherine and Marc Bolan. 2004. "Residential Trajectories Using Optimal Alignment to Reveal The Structure of Residential Mobility." *Sociological Methods & Research* 32(4):559-598.
- Struffolino, Emanuela, Matthias Studer, and Anette Eva Fasang. 2016. "Gender, Education, and Family Life Courses in East and West Germany: Insights from New Sequence Analysis Techniques." *Advances in Life Course Research* 29:66-79.
- Studer, Matthias, Gilbert Ritschard, Alexis Gabadinho, and Nicolas S. Muller. 2011. "Discrepancy Analysis of State Sequences." *Sociological Methods & Research* 40(3):471-510.
- Studer, Matthias. 2013. "WeightedCluster Library Manual: A Practical Guide to Creating Typologies of Trajectories in the Social Sciences with R." *LIVES Working Papers, NCCR LIVES* 24.
- Studer, Matthias. 2018. "Divisive Property-Based and Fuzzy Clustering for Sequence Analysis." in *Sequence Analysis and Related Approaches. Innovative Methods and Applications*, edited by G. Ritschard and M. Studer. Springer.
- Studer, Matthias and Gilbert Ritschard. 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.

- Van Winkle, Zachary and Anette Eva Fasang. 2017. "Complexity in Employment Life Courses in Europe in the Twentieth Century—Large Cross-National Differences but Little Change across Birth Cohorts." *Social Forces* 96(1):1-30.
- Warren, John Robert, Liying Luo, Andrew Halpern-Manners, James M. Raymo, and Alberto Palloni. 2015. "Do Different Methods for Modeling Age-Graded Trajectories Yield Consistent and Valid Results? 1." *American Journal of Sociology* 120(6):1809-1856.
- Wu, Lawrence L. 2000. "Some Comments on 'Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect.'" *Sociological Methods & Research* 29(1):41-64.

Figures

Figure 1. Quality of partitions with a different number of clusters, measured referring to different criteria. To ease comparison all the criteria were centered and standardized. Optimal solutions are those corresponding to maximal values, with the exception of the Hubert's C coefficient that should be minimized. Source: Multi-purpose Survey on Household and Social Subjects 2009.

Figure 2. Amount of deviations in partitions with a different number of clusters. For each cluster the figure reports: the average dissimilarity from sequences to cluster's medoid, the average dissimilarity from cases to other sequences in their clusters, the % of sequences with a standardized dissimilarity from the medoid higher than 1.5, the % of sequences with a standardized dissimilarity from other cases higher than 1.5, the % of core points, and the % of noise points. Source: Multi-purpose Survey on Household and Social Subjects 2009.

Figure 3. Plot of sequences and of cluster membership for partitions with a number of clusters between 4 and 9. In the top-plot sequences are arranged according to their membership to clusters in the 9-clusters solution, and within each cluster they are arranged based on the TSP seriation algorithm. The bottom-plot reports for each sequence the allocation of sequences to the clusters of the considered partitions (different colors correspond to different clusters within each partition). Source: Multi-purpose Survey on Household and Social Subjects 2009.

Figure 4. Flagged index plots for the 4-clusters partition extracted using PAM. For each cluster (row), the plot on the left reports cases flagged as regular by all the criteria used to identify deviant cases. Subsequent plots report cases flagged as deviant using alternative criteria: (a) sequences with an average dissimilarity from the others (in their cluster) higher than 1.5 times the general average (b) sequences with a dissimilarity from their cluster's medoid higher than 1.5 times the general average (c) noise points identified based on $\bar{d}_{0.95}^{NN5}$. In each plot, cases are arranged on the horizontal axis based on the TSP seriation algorithm. Source: Multi-purpose Survey on Household and Social Subjects 2009.

Figure 5. Flagged index plots for the 9-clusters partition extracted using PAM. For each cluster (row), the plot on the left reports cases flagged as regular by all the criteria used to identify deviant cases. Subsequent plots report cases flagged as deviant using alternative criteria: (a) sequences with an average dissimilarity from the others (in their cluster) higher than 1.5 times the general average (b) sequences with a dissimilarity from their cluster's medoid higher than 1.5 times the general average (c) noise points identified based on $\bar{d}_{0.95}^{NN5}$. In each plot, cases are arranged on the horizontal axis based on the TSP seriation algorithm. Source: Multi-purpose Survey on Household and Social Subjects 2009.

Figure 6. Confidence intervals for the coefficients of a multinomial logistic regression model for cluster membership (reference category: Cluster 2). Regressions are applied to the entire sample and to the sub-samples obtained by filtering out different types of deviant sequences: (a) sequences dissimilar from their cluster – i.e. sequences with an average dissimilarity from the others (in their cluster) or from their cluster's medoid higher than 1.5 times the corresponding general average (b) noise sequences identified based on $\bar{d}_{0.95}^{NN5}$ (c) both sequences dissimilar from their cluster and noise sequence. Source: Multi-purpose Survey on Household and Social Subjects 2009.

Legend for covariates: Int. = intercept; Birth cohort: 1959-1964 (reference), 1965-1969 (C.65-9), 1970-1975 (C.70-5); Macro-area of residence at the age of 15: North-East (R.NE), North-West (R.NW), Center (reference), South (R.S), and Islands (R.I); Highest parental education level: primary (reference), lower secondary (P.Edu2), upper secondary (P.Edu3), tertiary (P.Edu4); Working status of the mother when the respondent was 15: employed (M.Occ), not employed (reference).

APPENDICES

A. Additional Tables and Figures

Table A1. Distribution of the covariates included in the multinomial logistic regression models

	%
<i>Cohort</i>	
1959-1964	37.6
1965-1969	32.7
1970-1974	29.7
<i>Macro-area of residence</i>	
Nort-East	21.6
North-West	21.9
Center	17.1
South	28.9
Islands	10.5
<i>Highest parental education level</i>	
Primary	58.8
Lower Secondary	22.3
Upper Secondary	14.9
Tertiary	4.0
<i>Working status of the mother when the respondent was 15</i>	
Not working	61.3
Working	38.7
<i>N.</i>	<i>4,323</i>

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

Figure A1. Split-plots of the 4-clusters partition as the number of clusters increases. For each cluster the split-index plots report the sequences removed from the cluster when a higher number of clusters is extracted. In each plot, cases are arranged on the horizontal axis based on the TSP seriation algorithm. Source: Multi-purpose Survey on Household and Social Subjects 2009.

Note: For each cluster (on the rows) the index plot on the left reports those sequences that remain clustered together as the number of clusters increases from 4 to 9. The subsequent plots report cases that are assigned to other clusters as the number of clusters increases. For example, for cluster 1, the “Split5” plot shows that moving from 4 to 5 clusters, sequences characterized by a relatively longer period in Education are removed by cluster 1. The “Split6” plots identify sequences dominated by full-time temporary work (FT-Temp), which were not removed from the original clusters when the 5-clusters solution was obtained, but are instead removed when 6 clusters are formed. Note that when non-hierarchical clustering algorithms are used, these plots might be difficult to interpret, because the split plots cannot account simultaneously for the possible rearrangement of cases into clusters when the number of clusters changes. It is therefore not possible to state exactly which cases in the split plots would be isolated in dedicated clusters and which would instead be assigned to one of the clusters in the original 4-clusters partition. This adaptation of the flagged split plots is therefore particularly suitable and interpretable to describe hierarchical agglomerations of cases. Even so, also when partitioning algorithms are used, the plot shed some light on the features common to cases that are removed from the initial 4-clusters partition when it is refined increasing the number of clusters.

Figure A2. Flagged index plots for the 9-clusters partition extracted using PAM. For each cluster (row), the plot on the left reports cases flagged as regular by all the criteria used to identify deviant cases. Subsequent plots report cases flagged as deviant based on their dissimilarity to the others (in their cluster) using alternative thresholds: higher than (a) 1.5, (b) 2, and (c) 2.5 times the general average. In each plot, cases are arranged on the horizontal axis based on the TSP seriation algorithm. Source: Multi-purpose Survey on Household and Social Subjects 2009.

Figure A3. Flagged index plots for the 9-clusters partition extracted using PAM. For each cluster (row), the plot on the left reports cases flagged as regular by all the criteria used to identify deviant cases. Subsequent plots report cases flagged as deviant based on their dissimilarity to their cluster's medoid using alternative thresholds: higher than (a) 1.5, (b) 2, and (c) 2.5 times the general average. In each plot, cases are arranged on the horizontal axis based on the TSP seriation algorithm. Source: Multi-purpose Survey on Household and Social Subjects 2009.

Figure A4. Flagged index plots for the 9-clusters partition extracted using PAM. For each cluster (row), the plot on the left reports cases flagged as regular by all the criteria used to identify deviant cases. Subsequent plots report cases flagged as noise referring to alternative thresholds: (a) $\bar{d}_{0.90}^{NN5}$, (b) $\bar{d}_{0.95}^{NN5}$. In each plot, cases are arranged on the horizontal axis based on the TSP seriation algorithm. Source: Multi-purpose Survey on Household and Social Subjects 2009.

B. Alternative Strategies to Identify Deviating Sequences

We briefly review some possible alternative strategies that might apparently serve the purpose of identifying deviating sequences. Some authors (see e.g. Piccarreta and Studer, 2018) suggest to identify sequences weakly related to their cluster using their silhouette coefficient (Kaufman and Rousseeuw 2005). Nonetheless, the silhouette coefficient compares the average distance of a case to the cases in its cluster with the average distance to the cases placed in the closest cluster. Therefore, such index allows identifying sequences that lie between clusters, but it fails to identify cases that are relatively distant from their cluster without being close to other clusters.

Another strategy could be to apply a battery of clustering algorithms and to evaluate (for a given number of clusters) which cases would be grouped together by any algorithm and which would instead be aggregated to different cases depending on the adopted clustering method. While apparently promising, this approach suffers of the same limitations described for the silhouette coefficient. In fact, isolated sequences might be aggregated to the same cases irrespective of the clustering algorithm because they are too far from the rest of the data. Substantially, also in this case, it would be probably possible to identify cases lying between clusters but not necessarily cases in very peripheral positions.

Finally, we explored the possibility to flag deviating sequences using a classification tree (CART; Breiman et al. 1984) or a multinomial logistic regression to predict cluster membership based on the sequences' characteristics. The idea is to build an articulated set of variables describing sequences features – such as combinations of experienced states, presence of the different states, duration of the states, experienced transitions, time at transitions (Studer 2018) – and to predict

cluster membership based on them. Ideally, such approach should allow identifying the most relevant features of the sequences placed in each cluster. For misclassified sequences – with a predicted cluster different from the observed one – it would be theoretically possible to conclude that their characteristics are not coherent with those presented by the other sequences in their cluster. Nonetheless, the classification procedure would (at least theoretically) unveil the features common to the sequences in the same cluster, but would not necessarily prove useful to identify relevant types of deviating sequences. In addition, the application of this approach on our data showed that in many situations deviant cases were correctly classified despite very relevant structural differences with the other sequences in their cluster, because of marginal and almost irrelevant traits of similarity (for example similar initial tracks or similar ages at one specific events followed by very different unfolding of trajectories).