Article

Discrepancy Analysis of State Sequences

STUDER, Matthias, *et al*.

Abstract

In this article, the authors define a methodological framework for analyzing the relationship between state sequences and covariates. Inspired by the principles of analysis of variance, this approach looks at how the covariates explain the discrepancy of the sequences. The authors use the pairwise dissimilarities between sequences to determine the discrepancy, which makes it possible to develop a series of statistical significance–based analysis tools. They introduce generalized simple and multifactor discrepancy-based methods to test for differences between groups, a pseudo-R2 for measuring the strength of sequence-covariate associations, a generalized Levene statistic for testing differences in the within-group discrepancies, as well as tools and plots for studying the evolution of the differences along the time frame and a regression tree method for discovering the most significant discriminant covariates and their interactions. In addition, the authors extend all methods to account for case weights. The scope of the proposed methodological framework is illustrated using a real-world sequence data set.

Reference

STUDER, Matthias, *et al*. Discrepancy Analysis of State Sequences. *Sociological methods & research*, 2011, vol. 40, no. 3, p. 471-510

DOI : 10.1177/0049124111415372

Discrepancy Analysis of State Sequences

Matthias Studer, Gilbert Ritschard,
Alexis Gabadinho and Nicolas S. Müller
University of Geneva, Switzerland

Preprint of the article published in
Sociological Methods and Research. August 2011. Vol. 40(3), pp. 471-510.

Abstract

In this article we define a methodological framework for analyzing the relationship between state sequences and covariates. Inspired by the ANOVA principles, our approach looks at how the covariates explain the discrepancy of the sequences. We use the pairwise dissimilarities between sequences to determine the discrepancy which makes it then possible to develop a series of statistical-significance-based analysis tools. We introduce generalized simple and multi-factor discrepancy-based methods to test for differences between groups, a pseudo R^2 for measuring the strength of sequence–covariate associations, a generalized Levene statistic for testing differences in the within-group discrepancies, as well as tools and plots for studying the evolution of the differences along the timeframe and a regression tree method for discovering the most significant discriminant covariates and their interactions. In addition, we extend all methods to account for case weights. The scope of the proposed methodological framework is illustrated using a real-world sequence dataset.

Keywords: distance, dissimilarities, analysis of variance, regression tree, tree structured ANOVA, state sequence, optimal matching, homogeneity in discrepancies, Levene test, permutation test.

Author's note: Please address correspondence to Matthias Studer, Institute for Demographic and Life Course Studies, University of Geneva, Bvd Pont D'Arve 40, 1211 Geneva 4, Switzerland, Matthias.Studer@unige.ch.
All the numerical results and plots presented in this article can be reproduced with the `R` script provided as a supplemental file on the website of the journal.

1 Introduction

Optimal matching (OM) and, more generally, clustering of state sequences have become popular tools for analyzing life trajectories since their initial introduction in the social sciences in the late 1980s (Abbott and Forrest 1986; Abbott and Hrycak 1990). The popularity of these techniques is largely attributable to the holistic view they provide on the life course construction process. As emphasized by Billari (2005) such an approach considers the complete life sequence as a single unit of analysis, as opposed to event history analysis, for instance, which focuses on specific events of the life process such as marriage or the start of a new job.

In general, the OM approach consists of measuring the pairwise dissimilarities between sequences and then using the obtained values in an unsupervised clustering algorithm to build a typology of the observed sequences (i.e., to find homogeneous groups of sequences). Such analysis has proven to be an effective exploratory tool for discovering the main characteristics of a set of sequences without formulating any a-priori hypothesis. Further, it permits identification of typical trajectories and recurring structures in the sequences, thus bringing out fundamental descriptive information and making the data easier to comprehend. From a sociological point of view, groups determined by clustering techniques are useful for characterizing typical trajectories. Even more, it is often assumed that cases in a cluster all follow more or less an associated ideal type of trajectory. Clustering serves in this way for identifying the ideal types, which then in turn provide nice, simplified interpretations of the clusters.

Aside the understanding of the intrinsic characteristics of the sequences, in this article, we are interested in how the individual sequences are impacted by their context. A common practice in that perspective is to investigate the relationship of the identified sequence types—the clusters—with covariates such as sex and birth cohort. This is achieved, for example, by looking at the association between the cluster membership and the covariates, or by explaining cluster membership by means of logistic regressions or classification trees. The downside of this cluster-based approach, however, is that by representing the diversity of trajectories with a limited number of clusters, one inevitably loses the information about the diversity within each cluster. In the causal perspective, reducing the set of sequences to a limited number of standard trajectories is a too-crude approximation and would lead to considering deviations from the standard inside a cluster as non-explained error terms. Furthermore, knowledge of the cluster membership alone does not inform about the distances and differences between clusters. As a result, wrong conclusions may be drawn about the sequences–covariates relationships.

As a solution to this problem, we propose a set of methods—the discrepancy analysis—that allow to direct analysis of the sequence–covariate links (i.e., without any prior clustering). Our approach focuses on the discrepancy of the sequences, which we measure from their pairwise dissimilarities (OM distances, for example). This trick allows then a generalization of the ANOVA principles. The basic idea behind the ANOVA approach is to measure and test the part of the discrepancy among sequences that can be explained by covariates. For instance, one can assess what fraction of the differences between individuals' academic careers can be explained by sex or how the construction of one's familial life may differ according to social origin.

The article is organized as follows. In Section 2, we briefly review the relevant literature, and in Section 3 we describe the dataset on the transition from school to work used for illustrating the discussed methods. In Section 4 we introduce a dissimilarity-based measure of discrepancy for a set of possibly weighted sequences and discuss its interpretation. In Section 5 we propose methods to study the relationship between sequences and a single categorical variable; that is, we focus on the comparison of groups of sequences defined according to the levels of a given covariate. We derive a pseudo R^2 for measuring the share of sequence discrepancy explained by the grouping variable and introduce a pseudo F test for assessing the significance of the association. We extend both statistics to account for case weights. We propose also a Levene-like statistic for testing the homogeneity of within-group sequence discrepancies. To conclude the section, we illustrate the behavior of the statistics with simulations. In Section 6, we show how results can be further investigated and rendered to characterize the differences between groups. Section 7 deals with the multi-factor case for which we introduce original formulas that can account for case weights. Section 8 exploits the previous material in a tree-structured fashion to reveal the factors that best discriminate sequences. Finally, in Section 9, we provide a brief overview on how to apply the presented methods in R with our TraMineR package.

To avoid overloading the reader, detailed mathematical developments and discussions have been taken out of the main text and included in Appendices A to C.

2 Literature review

Several approaches have been used to study the association between objects described by their pairwise dissimilarities and a categorical covariate (Gower and Krzanowski 1999; Anderson 2001; McArdle and Anderson 2001; Mielke and Berry 2007; Cuadras 2008). Apart from the most popular ones, which generally adopt ANOVA/MANOVA principles, Mielke and Berry (1983, 2007) have suggested using ad hoc statistics based on sums of distances. Reiss, Stevens, Shehzad, Petkova, and Milham (2009) show that the two approaches are equivalent under certain conditions. The statistical significance of the association is generally assessed through permutation tests, although Mielke and Berry (2007) propose a parameterized approximation for the empirical distribution generated by the permutations. All of these authors consider only metric distances. In contrast, McArdle and Anderson (2001) extend the same approach to semi-metric dissimilarities. This latter solution can also be applied in a multi-factor case. As far as weights are concerned, an interesting contribution one finds in the literature is the paper of Delicado (2007), who accounts for weights with a test derived from Gower and Krzanowski's (1999) formulation for the single factor case.

The different methods are used in various domains such as genetics (Zapala and Schork 2006), the study of income density functions (Delicado 2007), or the study of neuroimaging data (Reiss et al. 2009). Gower and Krzanowski (1999) deal with situations where the number of dependent (response) variables is larger than the number of observations. In line with the work by McArdle and Anderson (2001), the main application field is ecology and especially the analysis of ecosystems by means of semi-metrics, such as that of Bray-Curtis. To the best of our knowledge, however, ANOVA-like tools have not yet been applied so far to life trajectory analysis.

The methods presented in this article are inspired from the work of Mielke and Berry (2007) and McArdle and Anderson (2001). We adapt their solutions for the study of state sequences in the social sciences and extend them, so that we can also account for case weights. We also complement the theoretical background of their approaches with notions such as the contribution to discrepancy derived from Batagelj's (1988) developments of the Ward criterion.

The analysis of differences in discrepancies between groups (i.e., the test of equality in within-group discrepancies) has rarely been considered for dissimilarity-based discrepancies. Anderson (2006) proposes two tests based on distances in an associated principal coordinate space. Studer, Ritschard, Gabadinho, and Müller (2010) consider a generalization of Bartlett's test that uses directly the original dissimilarities. The latter approach is not suitable, however, for weighted data.

Regarding tree-structured methods, there are only few recent attempts to apply them on objects characterized by their dissimilarities. Piccarreta and Billari (2007) propose a non-supervised dissimilarity-based divisive tree algorithm that grows the tree independently of any covariates. They applied it on sequence data. Similarly to the present paper, Geurts, Wehenkel, and d'Alché Buc (2006) propose the use of a kernel-based supervised method; however, their approach is limited to Euclidean distances. A more general dissimilarity-based method can be found in Piccarreta (2010). Unlike our proposal, her tree-growing method is not controlled by a statistical significance and therefore requires post-pruning. The tree algorithm considered in Section 8 is essentially the same as the one introduced by Studer, Ritschard, Gabadinho, and Müller (2009) and Studer et al. (2010). Here, we build on it and adapt it for weighted data.

3 Illustrative data set

Let us start with describing the application setting that will serve as an illustration throughout the article. We use data and the research question from McVicar and Anyadike-Danes' (2002) study of the school-to-work transition in Northern Ireland. The aim is to "identify the 'at-risk' young people at age 16 years and to characterize their post-school career trajectories" using information such as qualification at the end of compulsory schooling, family background and demographic characteristics. Table 1 presents the list of covariates available in the dataset. In addition, the data contains a "weight" variable for adjusting for response bias.

4 Discrepancy of a set of sequences

In this section, we define a measure of the discrepancy of a set of sequences. In a life course framework, the discrepancy measures the between-individual variability of the life trajectories. Therefore, higher discrepancy, for example, would reflect a greater level of uncertainty about the path followed by the individuals. Depending on the situations, such uncertainty may be interpreted either as a form of

Table 1. List of Covariates

Variable	Description	Values
sex	Gender	female, male
region	Location of school in North Ireland	Belfast, North Eastern, South Eastern, Southern, Western
religion	Religion	Catholic, Protestant
funemp	Father unemployed at time of survey	yes, no
fmpr	Father has a professional, managerial or related job	yes, no
livboth	Living with both parents at time of first sweep of survey	yes, no
grammar	Grammar school secondary eduction	yes, no
gcse5eq	Qualifications gained by the end of compulsory education: 5 or more GCSEs at grades A–C, or equivalent	yes, no

precariousness or, on the contrary, as a reflection of the multiplicity of choices the individuals face. The discrepancy concept considered here must be clearly distinguished from the within-individual longitudinal state diversity that can be measured with the longitudinal entropy (Widmer and Ritschard 2009), the turbulence (Elzinga 2010), or the complexity index (Gabadinho, Ritschard, Studer, and Müller 2010). The latter measure the diversity of states and transitions inside sequences, while the discrepancy assesses the diversity of the trajectories.

Aside from this diversity interpretation, the discrepancy is also the key concept for measuring the association between sequences and covariates. Decomposing it into explained between-groups and residual within-groups discrepancy permits measurement of the explained share of discrepancy and testing for differences between groups. These topics will be addressed in the next sections.

Since we cannot directly observe the distance to some "mean sequence", the discrepancy of sequences will be defined from their pairwise dissimilarities. There are many different ways of computing such dissimilarities, either through proximity measures counting common characteristics (Elzinga 2007) or using edit distances (Lesnard 2010). The most popular dissimilarity measure used for sequence analysis in the social sciences is the optimal matching (OM) edit distance, also known as generalized Levenshtein distance in computer science. It is defined as the lowest cost of transforming one sequence into the other by means of state insertions–deletions (indel) and state substitutions. The total transformation cost depends on the individual cost of each used operation. Those costs can be organized into an augmented substitution cost-matrix between states by considering each insertion–deletion as a substitution with a null element (i.e., by defining a row and a column for this null element).[1] The resulting OM distance satisfies surely the triangle inequality as long as the elements of this augmented substitution cost matrix verify it (Yujian and Bo 2007). When that is not the case, the resulting OM dissimilarity between two sequences x and y could be greater than the sum of their dissimilarities with some other sequence. Though we will use the OM distance with the costs defined in McVicar and Anyadike-Danes (2002) for our application example, the way of measuring the discrepancy described hereafter is in no way limited to the OM distance alone. Any other measure of dissimilarity between sequences could be used instead.

The presentation in the remainder of the section is based on the generalization of the Ward criterion made by Batagelj (1988). The concepts introduced may also be found in Anderson (2001), Reiss et al. (2009) and Mielke and Berry (2007), though these papers deal only with the unweighed case while we propose here formulas that account for weights as well.

4.1 Discrepancy based on dissimilarities

In the Euclidean case, the sum of squares SS—or inertia—may be expressed in terms of the pairwise squared Euclidean distances. Let $y = (y_i)$ be a vector of length n, w_i the weight associated to case i and W the total sum of weights. The sum of squares can be expressed as (see Appendix A):

$$SS = \sum_{i=1}^{n} w_i (y_i - \bar{y})^2 = \frac{1}{W} \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_i w_j d_{e,ij}^2 \quad (1)$$

where $d_{e,ij}$ is the Euclidean distance between i and j.

Following Mielke and Berry (2007), the concept of the sum of squares can be generalized to other dissimilarity measures by replacing the squared Euclidean distance $d_{e,ij}^2$ in the right hand side of Equation (1) with d_{ij}^{ν}, where d_{ij} is any possibly non-Euclidean measure of dissimilarity and ν a real positive

[1] Though this definition permits state dependent insertion–deletion costs, unique indel costs are most often used.

exponent, yielding:

$$SS = \frac{1}{W}\sum_{i=1}^{n}\sum_{j=i+1}^{n} w_i w_j d_{ij}^{\nu} \quad (2)$$

Although Mielke and Berry (2007) studied a wide range of values for the ν exponent, existing literature usually considers only either $\nu = 1$ or $\nu = 2$. We address the choice between these two values in Appendix B where we argue in favor of $\nu = 2$ for Euclidean metrics and $\nu = 1$ for non-Euclidean ones such as OM.

Applying the definition $s^2 = \frac{1}{W}SS$ of the sample variance, we get a fairly intuitive measure of the discrepancy of the sequence objects. Since the variance is theoretically defined for Euclidean distances, we prefer the term "discrepancy" for this more general setting. Interestingly, the discrepancy s^2 is equal to half of the weighted average of the pairwise dissimilarities; that is:

$$s^2 = \frac{1}{2W^2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j d_{ij}^{\nu} \quad (3)$$

4.2 Contribution to the sum of squares

Batagelj (1988) shows that the previous generalization of the sum of squares SS also implies that the dissimilarity $d_{x\tilde{g}}^{\nu}$ between a sequence x and the (possibly virtual) gravity center $\tilde{g}$ of a set G of sequences is (see Equation (21) in Appendix A):

$$d_{x\tilde{g}}^{\nu} = \frac{1}{W}\Big(\sum_{i=1}^{n} w_i d_{xi}^{\nu} - SS\Big) \quad (4)$$

According to Batagelj (1988), the notion of a gravity center holds for any kind of distances and objects, even though it is not clearly defined for complex non-numeric objects such as sequences. It is likely that the gravity center does not itself belong to the object space, exactly as the mean of integer values may be a real non-integer value.

Since $SS = \sum_x w_x d_{x\tilde{g}}^{\nu}$, each term $w_x d_{x\tilde{g}}^{\nu}$ under this summation may be interpreted as the contribution of x to the total sum of squares. Even though the gravity center may not be observable, Equation (4) provides a comprehensive way to compute the most central sequence—the medoid—of a set using weights. Searching the x that minimizes (4) is equivalent to minimizing the sum of the weighted distances from x to all other sequences. The same solution was, for instance, considered in the unweighed setting by Abbott (1990) for finding a representative sequence.

The non-negativity of this contribution automatically results when d^{ν} satisfies the triangle inequality (see Appendix A), while negative contributions to the discrepancy can occur when the triangle inequality does not hold. For non-Euclidean dissimilarities such as OM, it is therefore preferable to proceed with $\nu = 1$, which ensures the triangle inequality, rather than with squared dissimilarities (see Appendix B).

5 Comparing groups of sequences

Aside from evaluating the variability of a set of sequences, measuring discrepancy from pairwise dissimilarities permits generalization of the analysis of variance (ANOVA) principles to any dissimilarity measure. It allows computation of the share of discrepancy "explained" by a covariate and thus evaluation of the strength of the association between trajectories and a covariate. Although classical tests based on normality assumptions are not applicable in this case, the significance of the relation can be assessed through permutation tests, as discussed in Anderson (2001).

In Section 5.3, we introduce a new test to compare group discrepancy. In some situations, it may be of interest to test whether the discrepancies within groups differ significantly. We then discuss the interpretation and the visualization of the difference between state sequences. In the last subsection, we provide empirical insights on the behavior of the proposed tests with simulations.

5.1 Measuring association

When generalizing the notion of sum of squares to non-Euclidean measures of dissimilarity, the Huygens theorem (Equation 5) that states that the total sum of squares (SS_T) is the between sum of squares (SS_B) plus the residual within sum of squares (SS_W) remains valid (Batagelj 1988).

$$SS_T = SS_B + SS_W \quad (5)$$

Thus, we can apply the ANOVA machinery to sequence objects.

All terms in Equation (5) can be derived from formula (2). The total sum of squares (SS_T) and the within sum of squares (SS_W) are computed directly with formula (2), SS_W being simply the sum of the within sums of squares of each subgroup. The between sum of squares SS_B is then obtained by taking the difference between SS_T and SS_W. Using Equation (5), we can assess the share of discrepancy explained by a categorical or discretized continuous variable. In the spirit of ANOVA, this reduction of discrepancy is due to a difference in the positioning of the gravity centers $\tilde{g}_k$ of the classes k (Batagelj 1988). Hence, conceptually, we look for the part of the discrepancy that is explained by differences in group positioning, and we measure it with the R^2 formula (6). Alternatively, we may consider the F that compares the explained discrepancy to the residual discrepancy. The F formula is provided in Equation (7), where m is the number of groups.

$$R^2 = \frac{SS_B}{SS_T} \tag{6}$$

$$F = \frac{SS_B/(m-1)}{SS_W/(W-m)} \tag{7}$$

5.2 Assessing statistical significance

The statistical significance of the association (i.e., of the explained part of the discrepancy) cannot be assessed with Fisher's F distribution as in classical ANOVA.[2] The F statistic (7) does not follow a Fisher distribution with sequence objects for which the normality assumption is hardly defendable. Therefore, we consider a permutation test (Anderson 2001; Manly 2007) which works as follows. At each step we change the group—the value of the covariates—assigned to each sequence by means of a randomly chosen permutation of the group membership vector. We thus get an F_{perm} value for each permutation. Repeating this operation R times, we obtain an empirical non-parametric distribution of F that characterizes its distribution under independence (i.e., assuming the sequences are assigned to the cases independently of the explanatory factors). From this distribution, we can assess the significance of the observed F_{obs} statistic by means of the proportion of F_{perm} that are higher than F_{obs}. It is generally admitted that 5,000 permutations should be used to assess a significance threshold of 1% and 1,000 for a threshold of 5% (see Manly 2007, and Appendix C).

The issue is how we can or even if we should adapt such permutation tests to account for weights. We propose three solutions:

1. Replicate cases a number of times corresponding to the weights before performing the permutation. This approach supposes that weights are integer counts.

2. Replace at each step the simple permutation with a random assignment of covariate profiles to the sequences using distributions defined by the weights.

3. Proceed with permutations ignoring weights and use them for computing the statistics for each permutation.

When the weights stand for counts of aggregated cases, we should restore individual cases by replicating the aggregated ones. By permuting aggregates only, we would miss possible permutations of cases within aggregated groups and therefore end up with a less powerful test. The second option is more or less equivalent but can be used with non-integer weights. Both of these techniques assume that weights reflect an aggregation of independently drawn cases. However, when weights do not result from aggregation but are intended to improve the sample representativeness, as it is the case in the example *mvad* data, it would not be correct to replicate cases. For example, a weight of 4 would not mean that 4 cases were drawn, and hence replicating it 4 times would incorrectly inflate the sample size.

Thus, the first and the second solutions should be used with counts reflecting aggregation. The third one should be applied in cases where weights are aimed at improving the sample representativeness.

Figure 1 shows the empirical density curve of the F_{perm} statistics obtained with 5,000 permutations of the values of the variable *livboth* ("living with both parents"). The observed F_{obs} statistic is equal to 2.49. The associated significance is 0.21 and corresponds to the red (grey) area in the plot. With 21% of the random F's greater than the F_{obs}, we *cannot* conclude that the trajectory of young people differs significantly with the values of the *livboth* covariate.

[2] We have also considered the use of the Brown-Forsythe F^* statistic to account for unequal group discrepancy (Brown and Forsythe 1974b). However, in our experiments results were always almost the same as with the traditional F. For sake of simplicity, we do not develop further this option.

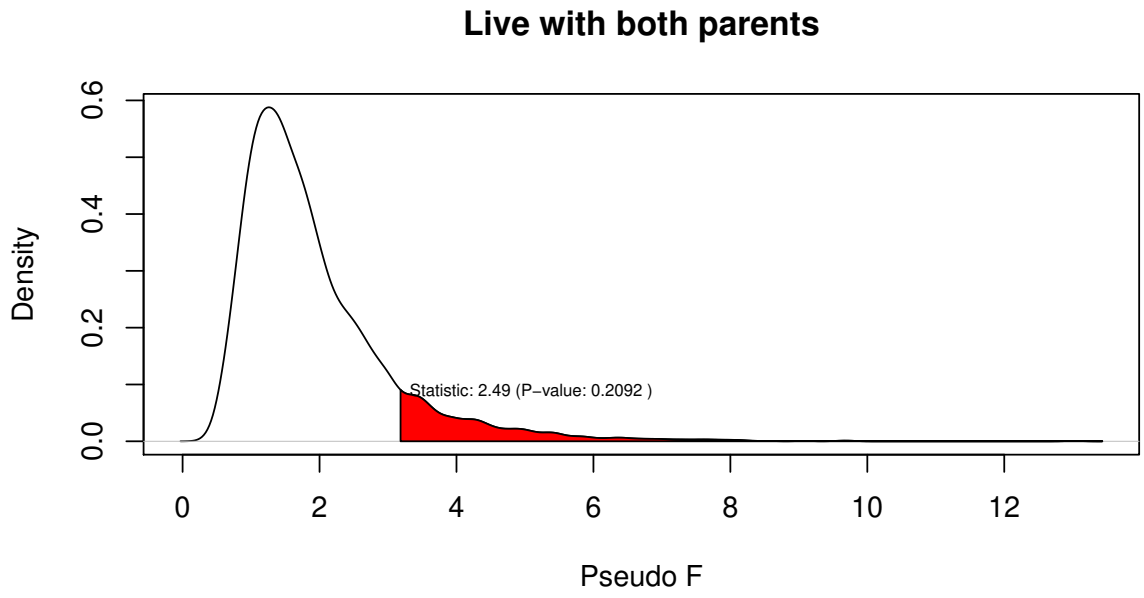

Figure 1. Empirical Distribution of the F Statistic under Independence with *livboth*

Table 2. Test of Association of Each Covariate with the School-to-Work Trajectories

	Non-squared dissimilarity			Squared dissimilarity			Non-squared, Unweighted		
	F	R^2	p-value	F	R^2	p-value	F	R^2	p-value
gcse5eq	104.09	0.128	0.000	184.09	0.206	0.000	67.69	0.087	0.000
grammar	59.34	0.077	0.000	89.87	0.112	0.000	23.16	0.032	0.000
fmpr	13.72	0.019	0.000	23.00	0.031	0.000	8.76	0.012	0.000
funemp	11.98	0.017	0.000	24.05	0.033	0.000	9.51	0.013	0.000
sex	11.03	0.015	0.000	13.85	0.019	0.001	6.84	0.010	0.000
region	5.44	0.030	0.000	7.26	0.039	0.001	5.50	0.030	0.000
livboth	2.49	0.003	0.211	2.61	0.004	0.240	2.23	0.003	0.033
religion	2.32	0.003	0.234	1.88	0.003	0.365	2.75	0.004	0.014

Table 2 summarizes the results of the discrepancy analysis using both squared and non-squared dissimilarities in the weighted case and non-squared dissimilarities only in the unweighted case. From the tests using weights, the trajectories significantly differ with qualification at the end of compulsory schooling (*gcse5eq*), type of compulsory schooling (*Grammar*), father employment status (*funemp* and *fmpr*), sex and region. We measured the strongest association for the end of compulsory schooling qualification, which lets us think that selection has occurred before the start of the sequences.

It is worth noting that results based on squared and non-squared dissimilarities are quite similar—both the ranking of covariates and the significance levels are the same—although the pseudo R^2 values are higher in the case of squared dissimilarities. This is a general result. The p-values associated with the unweighted test are lower, which may be explained by the additional variability that weights introduce in the estimation of the null F curve.

The significant effects found here agree with the significant effects on the cluster membership found by McVicar and Anyadike-Danes (2002), except for the *livboth* and *religion* covariates, which are not significant in our weighted ANOVA-like model. Since only variables with a significant effect in McVicar and Anyadike-Danes' (2002) study were included in the data set made available by those authors, it is not possible to find out covariates that significantly explain the sequence discrepancy while not affecting significantly the cluster membership.

5.3 Testing differences in within-group discrepancies

In some situations, it may be of interest to test whether the discrepancies within groups differ significantly. For instance, Elzinga and Liefbroer (2007) are interested in testing the de-standardization hypothesis stating that the family life trajectories of young adults become less similar over time (i.e., in testing for an increasing within-cohort discrepancy). However, their approach, which is to compare the mean pairwise similarities of four cohorts in 19 countries through 90% bootstrap confidence intervals, lends itself to criticism since by bootstrapping similarities instead of individuals it does not account for the strong correlation between similarities involving the same case. Better suited approaches are necessary to test differences in within-group discrepancies.

Studer et al. (2010) use a generalization of the Bartlett T (Bartlett 1937; Jobson 1991) for testing the homogeneity of the within-group discrepancies. Unfortunately, the value of the T Bartlett statistic is known to be very sensitive to the distribution of cases across groups (Manly 2007). It is therefore unsuited for randomizations of weighted data that modify this distribution at each draw.

We consider here an alternative approach based on the Levene test (Brown and Forsythe 1974a). The Levene statistic is known to be powerful with randomization tests (Manly 2007). From a geometric point of view, when testing homogeneity of discrepancy we are interested in measuring differences in the radius of the distribution of sequences within each group. Radii may be measured from the contributions to the within sums of squares and their general equality tested with an ANOVA procedure. Let $z_{i\ell} = d^{\nu}(x_i, \tilde{g}_\ell)$ be the dissimilarity between case i and the gravity center of group ℓ. The generalized Levene L statistic for testing group homogeneity is then the F statistic of this numeric variable. For the weighted case, it reads:

$$L = \frac{\sum_\ell w_\ell (\bar{z}_\ell - \bar{z})^2/(m-1)}{\sum_\ell \sum_i w_i (z_{i\ell} - \bar{z}_\ell)^2/(W-m)} \tag{8}$$

We propose again to evaluate the statistical significance of L through permutation tests. Though we test here differences in real values, namely the dissimilarities $z_{i\ell}$'s, we do not recommend comparing the observed L with the F distribution since those values are generally not normally distributed.

Table 3. Test of Homogeneity of the Within-group Discrepancies

	Non-squared dissimilarity		Squared dissimilarity		Non-squared, Unweighted	
	L	p-value	L	p-value	L	p-value
grammar	25.10	0.001	7.86	0.034	1.15	0.282
gcse5eq	14.15	0.005	7.94	0.021	0.57	0.432
funemp	7.88	0.030	6.00	0.054	7.50	0.008
religion	7.87	0.034	8.15	0.031	14.92	0.001
region	4.31	0.040	4.27	0.036	5.83	0.000
fmpr	5.57	0.075	3.52	0.156	0.02	0.885
livboth	0.88	0.487	1.13	0.439	0.94	0.331
sex	0.00	0.994	0.10	0.819	4.50	0.033

Table 3 provides the results of both tests for all considered covariates. As mentioned above, the Bartlett test is not reliable in the weighted case and is therefore shown only in the unweighted case. Using weighted data, we found, for instance, that young people who attended a grammar school have a discrepancy of their trajectories of 25.1, while those who did not have a discrepancy of 33.13. The difference is significative from our generalized Levene tests. In fact, grammar school opens the door to higher education trajectories, and most of the young Irish who attended a grammar school attempt to follow such a path, which results in a significantly lower discrepancy. We may also note that children of unemployed fathers show a significantly higher discrepancy and that there are significant differences according to the religion as well as across regions.

5.4 Simulation study of the tests' behavior

The aim of this section is to provide empirical insights on the behavior of the permutation tests of group differences. We study the tests on differences in both the general sequence pattern and the within discrepancy among sequences. We conducted a series of simulation studies to examine how the significance of the statistics evolves when we progressively change the characteristic sequence pattern of one of the groups. Three models were considered for generating artificial data sets.

1. Sequences with at most one transition and two possible states, with transition time—age at transition—generated with a normal model.

2. Sequences with at most one transition and two possible states, with transition time generated through a more realistic shifted log-logistic model.

3. Sequences with at most three transitions and eight possible states derived from the combination of three events occurring according to log-logistic models.

The first retained model depicts a simple situation where we can easily and independently control for the mean and variance of the age at which the transition occurs. The log-logistic model is more realistic for describing, for instance, the time to marriage or to a first childbirth. We retained a shifted version relevant for events that do not occur before a given age and studied the effect of varying once the start age—a positioning parameter—and once the log-logistic intensity—inverse of the scale parameter—for the second group. With the third set of simulations, we investigate sequences that may contain up to four states out of the eight resulting from different combinations of three successive possible events. For example, if we assume that the three events of interest are leaving home (L), marriage (M) and childbirth (C), and H stands for the initial state, the eight possible states would be H, L, M, C, LM, LC, MC, and LMC. In that example, M would mean married before leaving home and MC married with a child at parents' home. The occurrences of the events are each modeled with a different log-logistic distribution, and we vary only the parameters of one of them.

The normal model, though not very realistic for describing the hazard of a transition, permits independent examination of the effects of changes in position and in dispersion. The other models are more realistic. The log-logistic distribution is, for instance, used in parametric approaches of event history analysis (Blossfeld and Rohwer 2002). What makes it particularly interesting is that it allows for non-monotone risks. It is characterized by an intensity parameter λ and a shape parameter b. The inverse of λ is known as the scale parameter, and it is also the median of the distribution. Hence, an increase in λ reduces discrepancy but also changes the location. To control positioning independently of the discrepancy, we consider a start a parameter that specifies the threshold age where the log-logistic risk starts. The retained values for its b shape and λ intensity parameter are based on estimates obtained by Billari (2001b) in an analysis of age at first marriage in Italy. Finally, the last model considers multiple events and states that correspond typically to situations encountered in life course analysis as described above.

Table 4. Random Models Used for Generating the Simulated Data

	Normal	Log-logistic	Multiple Log-logistic
Transitions	$t \sim N(a, \sigma)$	$t \sim a + L(\lambda, b)$	$t_i \sim a_i + L(\lambda_i, b_i)$, $i = 1, 2, 3$
Parameters	a mean age σ age standard deviation	a start age λ intensity (inverse scale) b shape	a_1, a_2, a_3 start ages $\lambda_1, \lambda_2, \lambda_3$ intensities b_1, b_2, b_3 shapes
States	E_1 for $i < t$, E_2 for $i \geq t$	E_1 for $i < t$, E_2 for $i \geq t$	8 states from the combination of 3 events
Constant parts			$t_1 \sim 0 + L(0.078, 2.364)$ $t_3 \sim 20 + L(0.078, 2.364)$
Reference models			
for position change	$t \sim N(20, 4)$	$t \sim 0 + L(0.126, 2.364)$	$t_2 \sim 10 + L(0.126, 2.364)$
for scale change	$t \sim N(20, 4)$	$t \sim 0 + L(0.078, 2.364)$	$t_2 \sim 10 + L(0.078, 2.364)$
Variations	$20 \leq a \leq 25$ $4 \leq \sigma \leq 7$	$0 \leq a \leq 5$ $0.078 \leq \lambda \leq 0.205$	$10 \leq a_2 \leq 15$ $0.078 \leq \lambda_2 \leq .205$

The simulation design is as follows. For each of the three types of models, we start with the same set of parameters for the two groups and increase progressively in 20 steps one of the parameters for the second group while keeping the other parameters unchanged. At each step, we generate 1,000 artificial sets of sequences of length 40, with each set composed of $500\,\theta$ sequences for the reference group and $500(1-\theta)$ for the second group, with θ being an arbitrarily chosen proportion. For each artificial set, we compute the pairwise OM dissimilarities with a constant substitution cost of two and an indel cost of one. We then derive from them the values of the F and L statistics and their permutation test p-values, using once non-squared dissimilarities ($\nu = 1$) and once squared dissimilarities ($\nu = 2$).

The examined varying parameters are the mean age a for the normal model and the start age a for the log-logistic model. Those are location parameters. For changes in discrepancy, we varied the standard

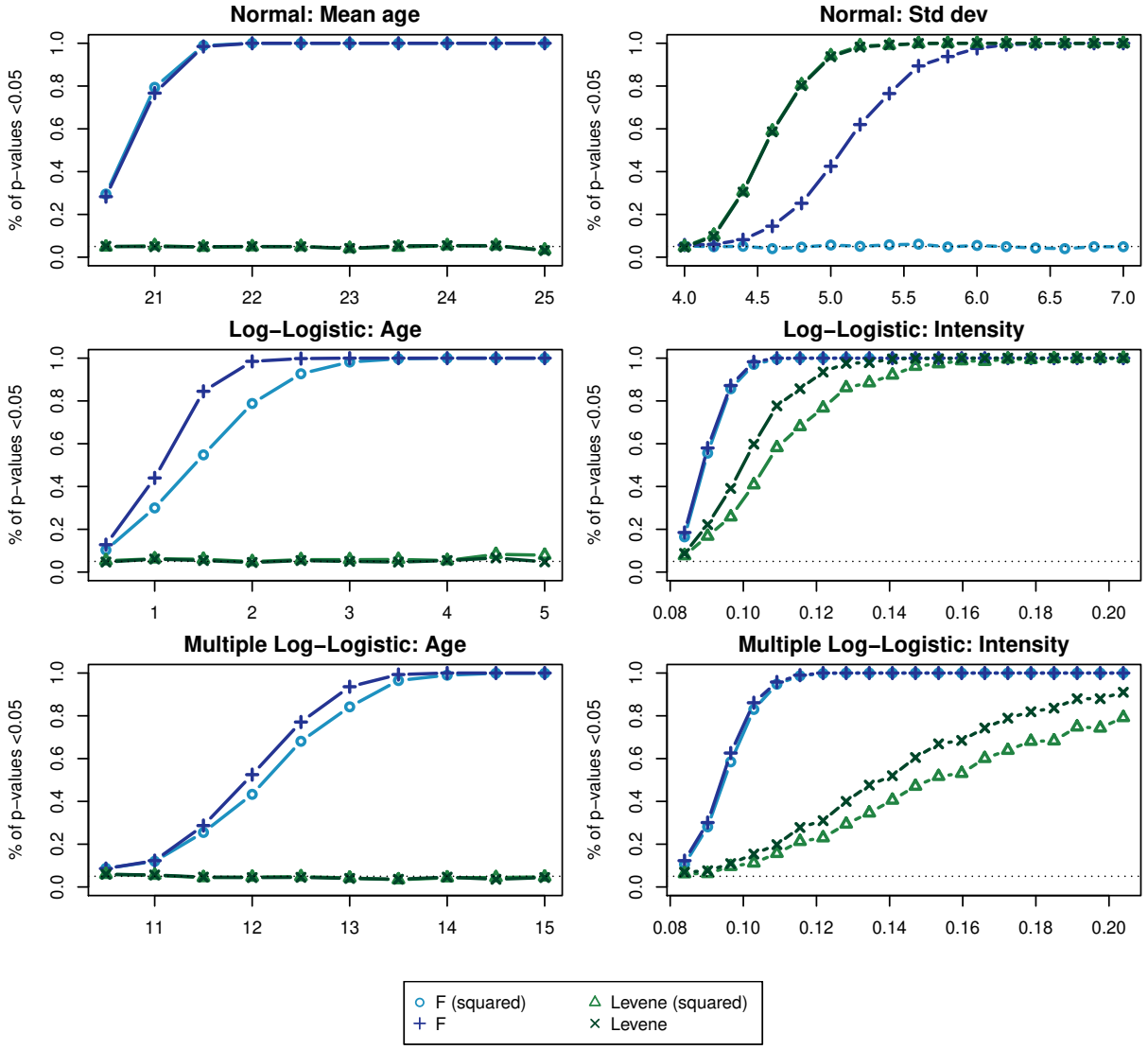

Figure 2. Simulation Results

deviation of the normal model and the intensity parameter of the log-logistic model. In the more complex case with multiple states generated from three log-logistics, we varied the parameters of one log-logistic distribution only, namely the one generating the second event.

Figure 2 summarizes the simulation results for a balanced distribution between the two groups (i.e., for $\theta = 50\%$).[3] Results reported are the percentage of p-values smaller than .05 at each of the stepped values of the considered varying parameter. Let us first look at the results for the F tests (blue curves). For the location parameter a (left plots), we get better significance when we use non-squared dissimilarities $\nu = 1$ than with $\nu = 2$ in the log-logistic case, while there are no clear differences between $\nu = 1$ and $\nu = 2$ in the normal case.

When we vary the scale parameter (i.e., the within-group discrepancy [right plots]), we observe that the F becomes significant with $\nu = 1$ in the normal case, which is a fallacious effect since we do not change location in that case. With $\nu = 2$, we get, as expected, non-significant F's. In the log-logistic case the F becomes also significant when λ increases, but it is not surprising here since λ also determines the position.

Let us now look at the L tests for the difference in within-group discrepancy. Unsurprisingly, when we vary the location parameter while maintaining the scales at the same level, the tests remain non-significant. When we vary the scale parameter, we observe similar significance of the L tests in the normal case while $\nu = 1$ dominates $\nu = 2$ in the log-logistic case.

[3]We ran the same simulations with proportions $\theta = 10\%$ and $\theta = 90\%$. We do not report the results here since they differ only marginally from those obtained with $\theta = 50\%$.

Results for the third set of simulations are similar to those of the single log-logistic, though differences between $\nu = 1$ and $\nu = 2$ are somewhat smoothed. This may be attributable to higher censoring of the concerned log-logistic that results from the occurrences of the other events.

Since data encountered in social sciences look more like those generated by the log-logistic models than like normal data, the use of $\nu = 1$ (i.e., non-squared dissimilarities) seems preferable in view of these simulation results. This goes in the direction of what we advocate in Appendix B on the basis that $\nu = 1$ guarantees the non-negativity of the contribution to the sum of squares when the dissimilarity satisfies triangle inequality. Altogether, we promote testing differences with $\nu = 1$ and confirming the conclusion with the squared version when the Levene test exhibits significant differences in within discrepancies.

6 Studying and rendering group differences

From the previous results, we learn that the trajectories of grammar school students not only differ significantly from the trajectories of students attending other schools, but also that the discrepancy of their trajectories is significantly lower. Nevertheless, these results by themselves do not tell us anything about how the trajectories differ among the different groups. To gain an idea of possibly existing differences, it is useful to display them visually. For example, index-plots (Scherer 2001) can be used, where each sequence is represented by a time line split into segments colored according to the corresponding occupational state. To account for weights, the line width can be adjusted according to the weight of the represented trajectory.

Figure 3 displays such weight-adjusted index-plots of grammar school and non-grammar school trajectories. Furthermore, to improve readability, we ordered the sequences according to the first dimension of a weighted principal coordinate analysis (PCO) (Gower 1966).[4] While ordering sequences by a principal coordinate facilitates the interpretation of the index-plot, the plots provide conversely useful information for interpreting the PCO axis. For instance, we observe in our case that the sequences are organized in a continuum ranging from higher education trajectories to training trajectories, while middle values correspond to employment-dominated trajectories. Comparing both populations, it appears that young people who attended grammar schools typically remain in the "school" state and are more likely to proceed to higher education, while those who attended other types of schools follow more diverse trajectories.

Figure 4 shows the evolution of the strength of association between the *Grammar* covariate and a sliding six months long sub-sequence of the trajectory. We choose a window length of six months which corresponds to a concrete horizon for most respondents. It has the effect of smoothing the curves while still rendering the main changes along time. The OM dissimilarity matrix was calculated for each six months windows and served for deriving the share of explained discrepancy and the value of the pseudo Levene statistic. Observing the evolution of these two statistics helps in identifying the periods over which the sequences differ the most. We observe that attending a grammar school has a long term effect with the strongest association appearing near the end of the studied trajectory. The curve of the Levene statistic indicates large differences in the discrepancy at the beginning of the sequences where most grammar students continue in the same school state while non-grammar students experience more diverse trajectories.

It is interesting also to look at how the within discrepancy evolves inside each group. Figure 5 depicts these evolution patterns using the same six-month sliding windows. It shows the discrepancy for each value of the *Grammar* variable and the overall discrepancy for comparison. We can see that the discrepancy of non-grammar school students gradually declines, which may be explained by the increasing number of youngsters who reach a final "employment" status. On the contrary, the discrepancy among grammar school students peaks after two years when some of them switch to higher education while others enter the labor market. We also observe that the differences of within-group discrepancies diminish over time, which is in accordance with the evolution of the Levene statistics presented above.

Depicting the evolution of the discrepancy over time can be seen as an alternative to studying the sequence of transversal entropies measuring state diversity at each time position (Billari 2001a; Widmer and Ritschard 2009). The latter approach is an analysis with windows of length one. Moreover, it can be shown that for such one-period windows the transversal Gini index—also known as quadratic entropy—is exactly the discrepancy derived from Hamming[5] distances with a unique substitution cost (Geurts et al. 2006). With lengthier windows, the solution proposed here permits accounting for discrepancies in both the sequencing and the temporality of the states.

[4] In order to be consistent with the rest of our analysis, we computed the PCO from the square roots of the dissimilarities since the PCO procedure automatically squares the provided dissimilarities. This amounts to set $\nu = 1$.

[5] The Hamming distance is OM without indels.

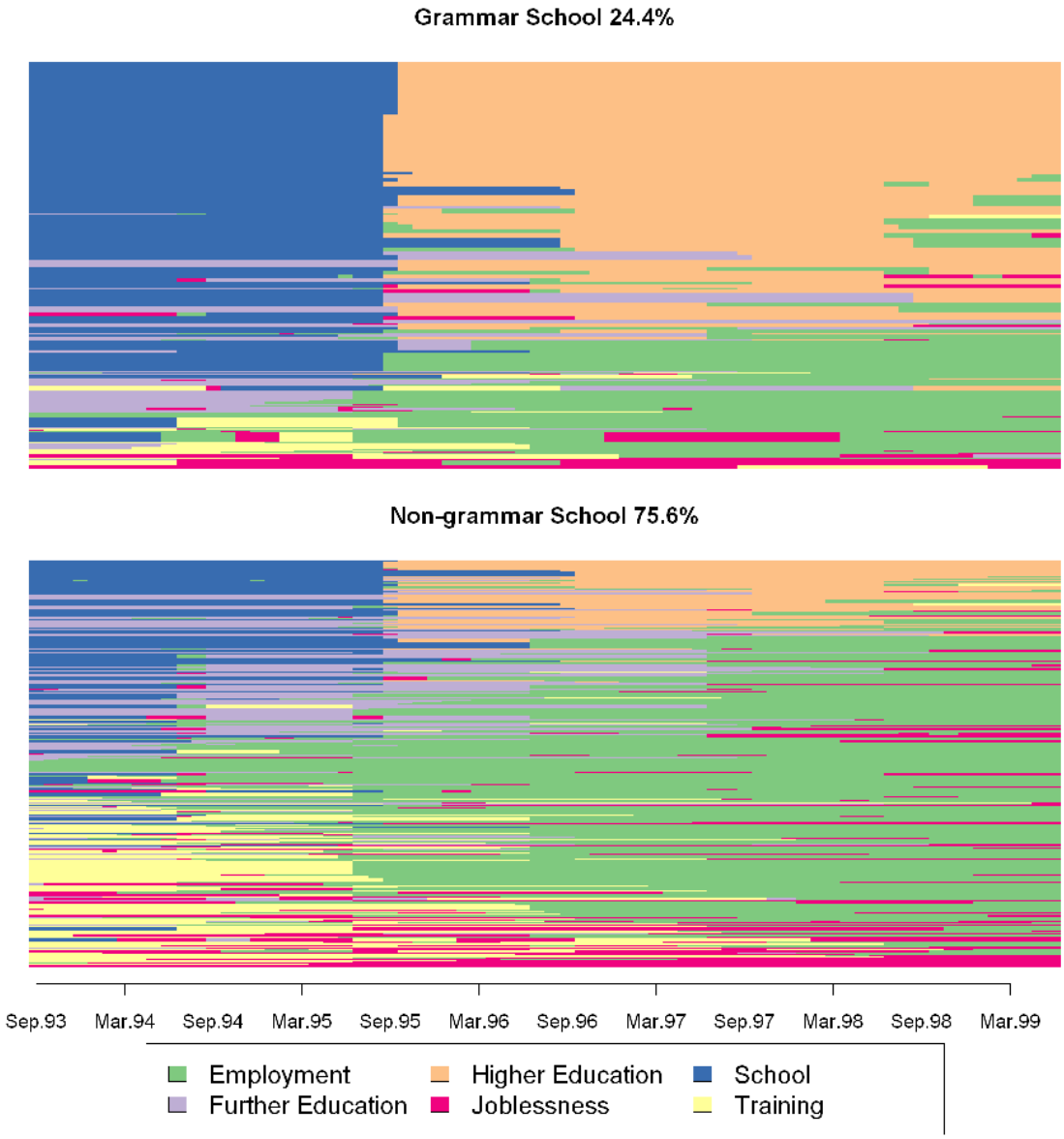

Figure 3. Trajectories of Grammar and Non-Grammar School Students

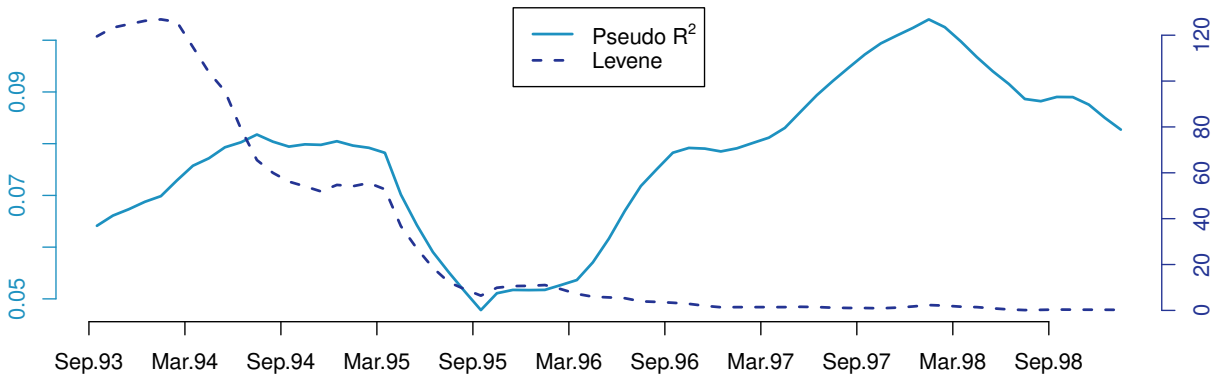

Figure 4. Time Evolution of the Pseudo R^2 and L, Six-months Sliding Windows

The proposed discrepancy analysis includes not only measuring the influence of a factor on the trajectories, but also depicting their diversity, both statistically and graphically. Conceptually, it aims

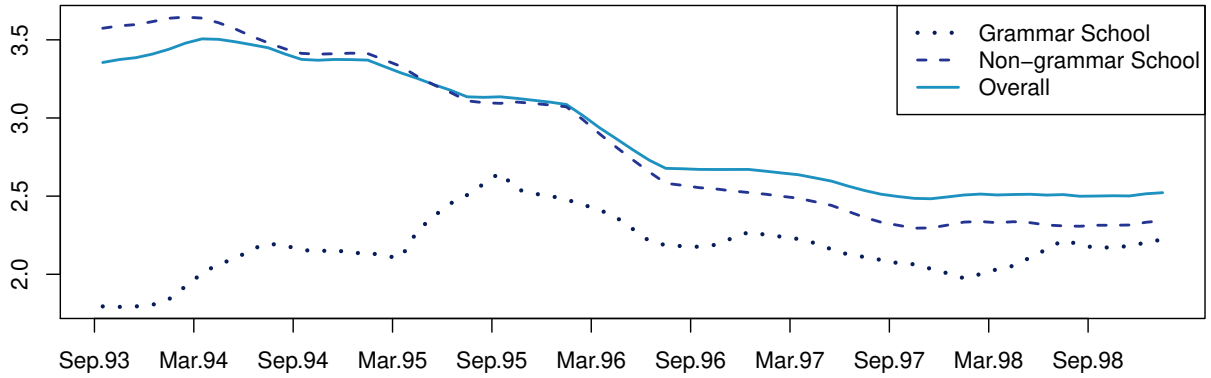

Figure 5. Time Evolution of Within-group and Overall Discrepancies, Six-months Sliding Windows

at depicting the influence of structures on the trajectory construction while assessing the ability for auto-determination within the structure.

Such an explanatory methodological framework based on the notion of discrepancy adheres particularly well to the life course paradigm. Indeed, in his formalization of this paradigm, Elder (1999) highlights the importance of studying the socio-historical context of the individuals as well as their ability to make their own choice within that context. This is precisely what discrepancy analysis does. It allows for studying the links between individuals' trajectories and their contexts, while at the same time preserving the notion of between-individual variability.

7 Multi-factor discrepancy analysis

In Section 5 we examined how to measure the bivariate association between the trajectory and each of the covariates considered independently. We consider here the generalization to the multi-factor case and, following the work of McArdle and Anderson (2001), adopt the framework of the general multivariate analysis of variance .

Formally, let $\mathbf{Y}$ be the $n \times q$ matrix with n observed values of q centered variables. In his work on principal coordinate analysis (PCO), Gower (1966, 1982) showed that the outer product $\mathbf{YY}'$, which has the sum of squares on its diagonal, can be expressed in terms of the matrix of pairwise squared Euclidean distances. McArdle and Anderson (2001) derived their generalized multi-factor MANOVA by combining this result with a multivariate linear model and replacing the Euclidean distances with dissimilarities. Here, we extend their proposition to account for weights.

Let $\mathbf{X}$ be the $n \times m$ matrix with the values of m predictors—contrasts for coding M factors—including a first column consisting of ones for the constant and $\hat{\mathbf{Y}}$ the matrix of values predicted from $\mathbf{X}$ with the linear model. The sum of total weighted sums of squares (SS_T) may be partitioned into predicted (SS_B) and residual (SS_R) weighted sums of squares.

$$tr\big(\mathbf{Y}'\mathbf{WY}\big) = tr\big(\hat{\mathbf{Y}}'\mathbf{W}\hat{\mathbf{Y}}\big) + tr\big(\mathbf{R}'\mathbf{WR}\big) \tag{9}$$

where $\hat{\mathbf{Y}}$ is the matrix of the $\mathbf{Y}$ values predicted by $\mathbf{X}$, $\mathbf{R}$ the matrix of residuals, and $\mathbf{W}$ the weight diagonal matrix. According to the weighted linear regression model, we have $\mathbf{W}^{\frac{1}{2}}\hat{\mathbf{Y}} = \mathbf{H}\mathbf{W}^{\frac{1}{2}}\mathbf{Y}$ and $\mathbf{W}^{\frac{1}{2}}\mathbf{R} = (\mathbf{I}-\mathbf{H})\mathbf{W}^{\frac{1}{2}}\mathbf{Y}$, where $\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\mathbf{WX})^{-1}\mathbf{X}'\mathbf{W}^{\frac{1}{2}}$ is the symmetric idempotent "hat" matrix that is adapted for the weighted case here. Since $\mathbf{W} = \mathbf{W}^{\frac{1}{2}}\mathbf{W}^{\frac{1}{2}}$, using the property $tr\big(\mathbf{AB}\big) = tr\big(\mathbf{BA}\big)$ for conformable matrices, Equation (9) may be rewritten as:

$$tr\big(\mathbf{WYY}'\big) = tr\big(\mathbf{W}^{\frac{1}{2}}\mathbf{H}\mathbf{W}^{\frac{1}{2}}\mathbf{YY}'\big) + tr\big[\mathbf{W}^{\frac{1}{2}}(\mathbf{I}-\mathbf{H})\mathbf{W}^{\frac{1}{2}}\mathbf{YY}'\big] \tag{10}$$

Let us now look at Gower's result that expresses $\mathbf{G} = \mathbf{YY}'$ in terms of the pairwise squared distances. In the formulation retained by McArdle and Anderson (2001), we have $\mathbf{G} = -\frac{1}{2}\big(\mathbf{I} - \frac{1}{n}\mathbf{11}'\big)\mathbf{D}\big(\mathbf{I} - \frac{1}{n}\mathbf{11}'\big)$, where $\mathbf{1}$ is a vector of ones of length n, $\mathbf{I}$ the identity matrix, and $\mathbf{D}$ the $n \times n$ matrix of the squared pairwise Euclidean distances. Here, we have to adapt this definition for $\mathbf{y}$ variables centered on their weighted means. In that case, it reads as follows:

$$\mathbf{G} = -\frac{1}{2}\Big(\mathbf{I} - \frac{1}{W}\mathbf{11}'\mathbf{W}\Big)\mathbf{D}\Big(\mathbf{I} - \frac{1}{W}\mathbf{W11}'\Big) \tag{11}$$

The generic element of $\mathbf{G}$ is $g_{ij} = -\frac{1}{2}\big(d_{ij}^2 - \bar{d}_i^2 - \bar{d}_j^2 + \bar{d}^2\big)$, where $\bar{d}_i^2$, $\bar{d}_j^2$ and $\bar{d}^2$ are respectively the row i, the column j, and the overall weighted average of the squared Euclidean distances. It can be shown that when $q = 1$, each diagonal element g_{ii} of $\mathbf{G}$ is the contribution of the i-th individual to the sum of squares as defined in Equation (4).[6]

Setting $\mathbf{H_W} = \mathbf{W}^{\frac{1}{2}}\mathbf{H}\mathbf{W}^{\frac{1}{2}}$, the total SS_T, between SS_B and within SS_W sums of squares of interest may be rewritten in terms of this adapted $\mathbf{G}$ matrix as:

$$SS_T = tr\big(\mathbf{WG}\big) \tag{12}$$

$$SS_B = tr\big(\mathbf{H_W G}\big) \tag{13}$$

$$SS_W = tr\big[(\mathbf{W} - \mathbf{H_W})\mathbf{G}\big)\big] \tag{14}$$

The idea is to substitute the pairwise dissimilarities d_{ij}^{ν} for the squared Euclidean distances that define $\mathbf{D}$ in Equation (11). Assuming such a substitution and using formulas (12–14), we can derive global pseudo-R^2 and pseudo-F statistics as defined in Equations (6–7). Instead of the number of groups, m should be set here to the number of columns of $\mathbf{X}$ (i.e., to the total number of contrast and/or indicator variables necessary for coding the M factors).

For $M = 1$ (i.e., in the case of a single factor), computing the SS_T and SS_W with the formula (2 on page 5) as shown in Section 5 gives exactly the same results as the matrix formulation considered here. However, the direct computation of the sums of squares is about 10 times faster.

We may also consider the contribution of each covariate to the total discrepancy reduction. As with multi-factor ANOVA, there are different ways of looking at these individual contributions. Shaw and Mitchell-Olds (1993) distinguish, among others, between two methods called Type I and Type II, respectively. The Type I method is incremental, which means that covariates are successively added to the model and the contribution of each covariate is measured by the SS_B increase that results when it is introduced. With this method, the measured impact of each covariate depends on the order in which the covariates are introduced. With the Type II method, known to be robust in the absence of interaction effects, the contribution of each covariate is measured by the reduction of SS_B that occurs when we drop it out from the full model (i.e., from the model with all covariates). We retain the second method and hence compute the following F for each covariate v

$$F_v = \frac{(SS_{B_c} - SS_{B_v})/p}{SS_{W_c}/(W - m - 1)} \tag{15}$$

where the SS_{B_c} and SS_{W_c} are the explained and residual sums of squares of the full model, SS_{B_v} the explained sum of squares of the model after removing variable v, and p the number of indicators or contrasts used to encode the covariate v.

As in the single discrepancy analysis, the F distribution is not relevant for the pseudo-F, and we consider again permutation tests for assessing the significance of the F statistic. Since F_v is intended for testing the conditional independence of v, its null distribution is obtained by permuting only the covariate v while the global F statistic is computed by permuting the whole profiles. Thus, for a complete multi-factor analysis with profiles defined by M factors, $1 + M$ permutation tests are required, which may be quite time-consuming.

Table 5. Multi-Factor Discrepancy Analysis

	Full Model			Backward Model		
Variable	F_v	ΔR_v^2	Sig	F_v	ΔR_v^2	Sig
gcse5eq	51.91	0.060	0.000	55.72	0.065	0.000
grammar	20.77	0.024	0.000	21.44	0.025	0.000
sex	5.47	0.006	0.002	5.30	0.006	0.003
funemp	3.59	0.004	0.039	3.83	0.004	0.028
fmpr	3.30	0.004	0.054			
region	3.19	0.015	0.004	3.37	0.016	0.003
religion	2.29	0.003	0.212			
livboth	1.80	0.002	0.405			
	F_{tot}	R_{tot}^2	Sig	F_{tot}	R_{tot}^2	Sig
Global	14.96	0.190	0.000	19.55	0.182	0.000

[6]Though it is not a concern here, this result can easily be extended for $q > 1$.

Let us look at what a multi-factor analysis gives for our illustrative example. Table 5 shows the results for two models: the complete model with all variables and a model obtained after removing non-significant covariates through a backward stepwise process. The tests were conducted using 5,000 permutations.

Both models provide overall significant information about the discrepancy of the trajectories since both global F statistics are significant. The full model explains a slightly higher part of the discrepancy ($R^2 = 0.190$) than does the backward model ($R^2 = 0.182$), but it contains non-significant covariates.

In the full model, the variable "qualification gained at the end of compulsory education" (*gcse5eq*) is the most significant covariate. If we remove this variable, the R^2 of the model (= 0.190) decreases by 0.060. The difference is significant since we have $F_{gcse5eq} = 51.91$, which was never attained with 5,000 permutations. As before, the variable *religion* is not significant. Removing it from the model reduces the R^2 by only 0.003 and results in a $F_{religion}$ value of 2.29 and a p-value of 0.208. Likewise, the variable "having a professional, managerial or related father" (*fmpr*) loses its significance in the multi-factor case. In fact, the variable *fmpr* becomes non-significant as soon as we control for "father's unemployment" (*funemp*), as the two variables are strongly correlated and "father's unemployment" is the most significant one.

The multi-factor approach provides information about the proper effect of the covariates on the occupational trajectory (i.e., about the part of the total effect that is not accounted for by factors that are already introduced). In that sense, the multi-factor approach is complementary to the single univariate discrepancy analysis which informs on the raw effect of each covariate. Nevertheless, while the multi-factor approach permits us to know which effects are significant, it does not tell us much about what the effects are (i.e., about how trajectories may change with the value of the covariates). To answer such questions, we propose a tree approach which can be seen as an extension of the graphical display shown in Figure 3.

8 Tree-structured analysis of sequences

In this section, we complement the sequence discrepancy analysis with the regression tree method introduced in Studer et al. (2009, 2010) which we extend to account for weighted sequences. Regression trees work as follows (Morgan and Sonquist 1963; Breiman, Friedman, Olshen, and Stone 1984). They start with all individuals grouped in an initial node. Then, they recursively partition each node using values of a predictor. At each node, the predictor and the split are chosen in such a way that the resulting child nodes differ as much as possible from one another or have, more or less equivalently, lowest within-group discrepancy. The process is repeated on each new node until a certain stopping criterion is reached.

The recursive partitioning generated by a tree is known to provide an easily comprehensible view of how each newly selected covariate nuances the effect of covariates introduced at earlier levels. This requires the display of relevant information about the distribution at each node. We could represent the medoid (i.e., the observed sequence that minimizes the dissimilarity (Equation 4) between the sequence and the group gravity center). It would be instructive to render the within-group discrepancy as well. Although this is not obvious for any kind of complex objects, displaying index-plots like those used in Figure 3 provides a good solution for state sequences. For a somewhat more synthetic view, we could also consider representative plots (Gabadinho, Ritschard, Studer, and Müller 2011b) that show the minimal set of sequences for each node that would be necessary to ensure a given coverage of the sequences at that node.

Beside the displayed node content, the originality of the tree-structured analysis of sequences resides in the use of a splitting criterion derived from the pairwise dissimilarities, namely the univariate pseudo-R^2 that we described in Section 5. At each node, we select thus the predictor and the binary split for which we get the highest pseudo-R^2 (i.e., the split that accounts for the greatest part of the object discrepancy). An alternative would be to use the significance of the univariate pseudo-F. However, since this significance must be determined through permutation tests, the time complexity would be excessive if we had to repeat it for each predictor and possible split. Therefore, we consider the F significance only as a stopping criterion; that is, we stop growing a branch as soon as we get a non-significant F for the selected split. In that way, permutations need to be run only once at each node, which remains tractable. The extension of this tree method for weighted cases is obtained by using the pseudo-R^2 formula and the F testing method that we propose in this article at each step of the tree-growing process.

Using the pseudo-R^2 as the splitting criterion inevitably means that we could only build binary trees. The R^2 does not penalize for the number of groups and would hence always select the maximum number of groups if we allowed n-ary splits. Using the R^2 adjusted for the number of groups, as it is used in multiple regressions, would not solve the problem since the adjusted R^2 is known to insufficiently penalize complexity. Using information criteria such as the BIC would also not be suitable, as such criteria are hardly derivable in a case where the distribution of the statistics (R^2, F or SS_W) under the independence

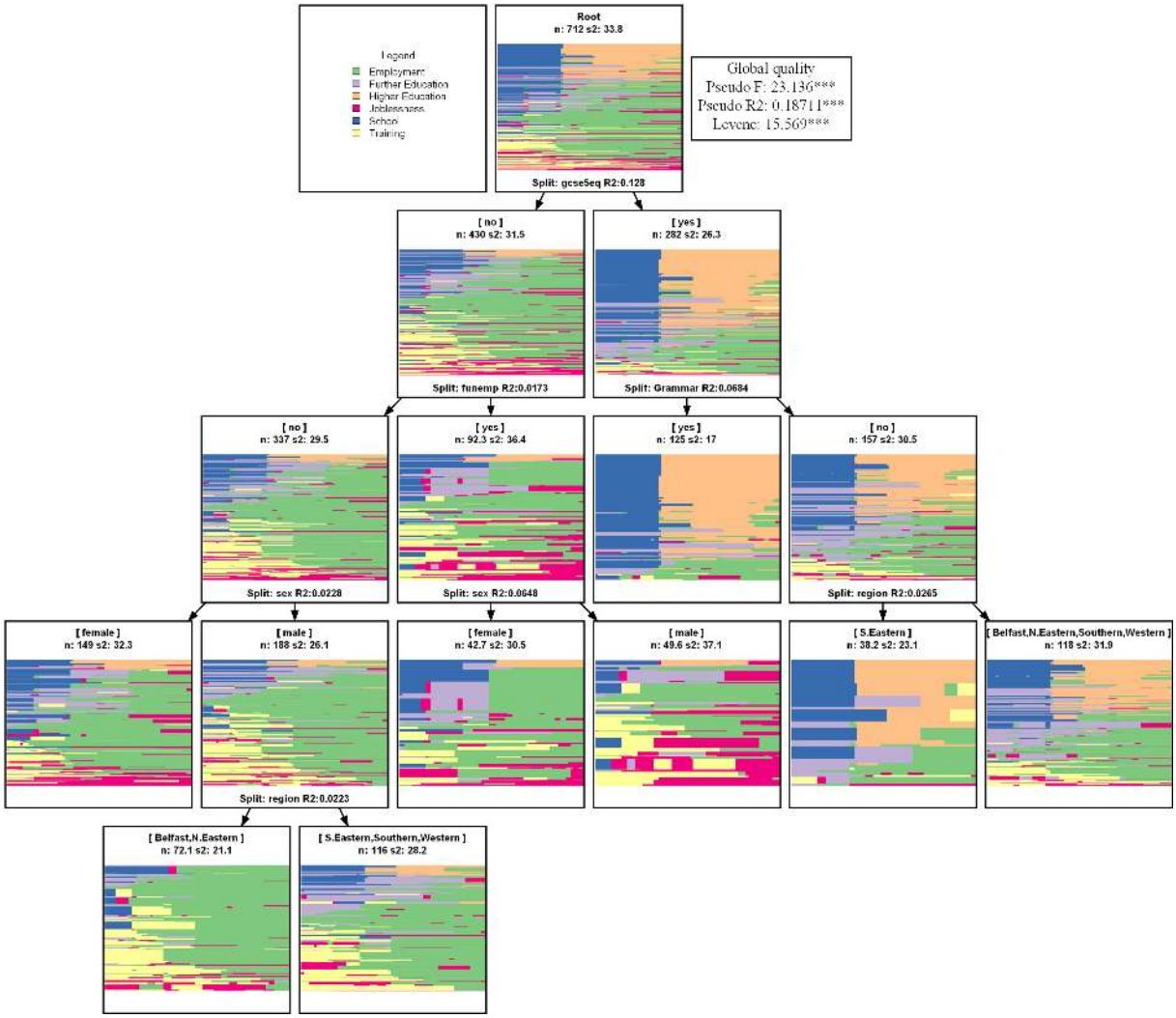

Figure 6. Sequence Regression Tree

hypothesis is not known.

The global quality of the tree can be assessed through the association strength between the sequences and the leaf (terminal node) membership. The global pseudo-F provides a way of testing the statistical significance of the obtained segmentation, while the global pseudo-R^2 provides a measure of the part of the total discrepancy that is explained by the tree.

Figure 6 shows the dissimilarity tree grown using our weighed example dataset. The chosen stopping criteria are a p-value of 5% for the F test, a minimal leaf size of 5% of the total sum of weights, and a maximal depth of 5. In each node, we see the plot of the individual sequences as well as the node size, the sum of weights, and the discrepancy s^2 within the node. At the bottom of each parent node, we indicate the retained split predictor with the associated R^2, while the definition of the binary split may be inferred from the indication at the top of the child nodes.

The overall R^2 of the tree is 0.187, which falls between the global R^2 of the full and backward models in Table 5. However, the results are now much easier to interpret. Moreover, the tree automatically accounts for interaction effects that were not considered in the multi-factor discrepancy analysis. We observe, for instance, that "attending grammar school" discriminates better among students who finished the compulsory schooling with high grades than among those who obtained lower grades. Likewise, we can see that having an unemployed father seems to affect primarily young male Irish with low grades at the end of compulsory schooling (*gcse5eq*).

9 Running sequence discrepancy analysis in R with TraMineR

We implemented the methods presented in this article into TraMineR (Gabadinho, Ritschard, Müller, and Studer 2011a), which is a free package for the R statistical environment (R Development Core Team

2008). Below, we briefly show how to run the discrepancy analysis features discussed here. We would also like to refer our readers to the TraMineR User's guide (Gabadinho, Ritschard, Studer, and Müller 2009), which provides a detailed overview of other features offered by the package, especially of the rendering of sequences and the computation of dissimilarities. Our readers can reproduce the results we present here, as the *mvad* dataset, which we use, has been made available as part of the TraMineR package, thanks to the authorization of McVicar and Anyadike-Danes.

To begin the analysis, we load the *mvad* data and create a weighted state sequence object using the commands below. The state variables from September 1993 to June 1999 are in columns 17 to 86 of the data frame.[7]

```
R> library(TraMineR)
R> data(mvad)
R> mvadseq <- seqdef(mvad[, 17:86], weights = mvad$weight)
```

Next, we compute the OM dissimilarity matrix with the indel and substitution costs used by McVicar and Anyadike-Danes (2002).

```
R> subm.custom <- matrix(
+     c(0,1,1,2,1,1,
+       1,0,1,2,1,2,
+       1,1,0,3,1,2,
+       2,2,3,0,3,1,
+       1,1,1,3,0,2,
+       1,2,2,1,2,0),
+     nrow = 6, ncol = 6, byrow = TRUE)
R> mvaddist <- seqdist(mvadseq, method = "OM", indel = 1.5, sm = subm.custom)
```

To perform the univariate discrepancy analysis and to test for homogeneity of discrepancy, we call the `dissassoc()` function which takes four arguments: the dissimilarity matrix, the factor (`group`), the number of permutations (`R=1000` by default), and an optional `weights` argument. The results presented in Section 5 were obtained with the following code:

```
R> dissassoc(mvaddist, group = mvad$gcse5eq, R = 5000,
+     weights = mvad$weight, weight.permutation="diss")
```

Likewise, we generated Figures 4 and 5 with `seqdiff()` as shown below.

```
R> Grammar.diff <- seqdiff(mvadseq, group = mvad$Grammar,
+     seqdist_arg = list(method = "OM", indel = 1.5, sm = subm.custom))
R> plot(Grammar.diff, stat = c("Pseudo R2", "Levene"))
R> plot(Grammar.diff, stat = "discrepancy")
```

The multi-factor results listed in Table 5 were obtained with the `dissmfac()` function. The model is specified as a classical R formula with the dissimilarity matrix on the left-hand side. We use the `data` argument to specify the *data.frame* containing the covariates.[8]

```
R> dissmfac(
+     mvaddist ~ gcse5eq + Grammar + funemp + catholic + male + fmpr + livboth + region,
+     data = mvad, R = 5000, weights = mvad$weight)
```

To carry out a tree-structured analysis of the sequences, we use the `seqtree()` function. The dissimilarity matrix and the predictors are passed to the function in the same way as in `dissmfac()`. Stopping criteria can be set with the arguments `minSize` for the minimum node size, `maxdepth` for the maximum tree depth and `pval` for the minimum required p-value. As for `dissassoc()`, the `R` argument controls the number of permutations for computing the p-values. Notice that it is not necessary to specify the weights since they are already attached to the state sequence `mvadseq` object.

```
R> mvadtree <- seqtree(
+     mvadseq ~ gcse5eq + Grammar + funemp + catholic + male + fmpr + livboth + region,
+     data = mvad, minSize = 30, maxdepth = 5,  R = 5000, pval = 0.01, diss = mvaddist)
R> print(mvadtree)
```

The `print()` command produces a text output of the tree. The tree can also be plotted with `seqtreedisplay()`. This function uses the free GraphViz software (Gansner and North 1999).[9] Hence, it must be installed and accessible for the function to work properly. The tree in Figure 6 was obtained with the following command.

```
R> seqtreedisplay(mvadtree, type = "I", sortv = cmdscale(sqrt(mvaddist), k = 1))
```

[7]For details on TraMineR functions such as `seqdef, seqdist, dissassoc` presented here see the reference manual or type for instance `?dissassoc` in the R console to access the on-line help.

[8]The `region` factor was built from the 5 region binary dummies in the *mvad* data frame with coding not shown here.

[9]The program can be downloaded from `http://www.graphviz.org/`.

10 Conclusion

In this article, we proposed a set of tools for analyzing the relationship between discrete sequences and one or more covariates. Besides the fact that the methods we propose are of interest for the analysis of state sequence, we believe that they also provide an innovative alternative to the traditional cluster-based sociological analysis of life trajectories.

The starting point of the new methodology introduced hereby is the definition of the discrepancy of the sequences in terms of their pairwise dissimilarities. Afterwards, the methods proceed with the transposition of the ANOVA concepts to this generalized discrepancy framework. They include single and multi-factor ANOVAs, the measure of the strength of sequence–covariate associations with pseudo R^2's, a generalized Levene test of equality of within-group discrepancies, tools and plots for investigating the evolution of the group differences along the timeframe, and a regression tree method for sequence data. Since normality of sequences is not defendable, the statistical significance of the proposed statistics is assessed through permutation tests. Up to this point, similar approaches have already been considered in the literature, but only for non-sequence complex objects such as ecosystems. In addition to the application on sequence data, the generalized Levene test and the procedure accounting for case weights in the measures and tests are the main original methodological contributions of this article.

As far as sociological analysis is concerned, the proposed methodology opens new perspectives besides the traditional cluster-based approach. In short, this cluster-based approach consists of associating each trajectory in a given set to some related ideal type. From a descriptive standpoint, this approach has proven to be effective in uncovering the underlying structure of a set of sequences, which makes the data easier to understand. However, relying on clusters for studying the relationship between sequences and their context can be criticized on the basis that reducing the set of sequences to a limited number of standard trajectories is a rather crude approximation and would lead to considering deviations from the standard inside a cluster as non-explained error terms. As a result of this approximation, wrong conclusions may be drawn about relationships between the sequences and their context. On the other hand, the approach we propose here takes into account explicitly how the individual characteristics affect the trajectory followed by each individual.

Furthermore, in this paper we adopted an explanatory methodological framework that complies with the life course paradigm (Elder 1999) by accounting for the individuals' ability to make their own choices within their socio-historical backgrounds when estimating the sequence–context relationship. By focusing on the discrepancy of the sequences, they allow studying the link between the trajectories and their context while preserving the notion of between-individual variability.

The choice of the measure of dissimilarity between sequences is a recurrent debate in the social sciences (Dijkstra and Taris 1995; Wu 2000; Elzinga 2003), which is beyond the scope of the present paper. Although we have illustrated the methods using an optimal matching edit distance, the methods considered in this paper are by no way limited to optimal matching. They work with any dissimilarity measure. Moreover, running the statistical tests with different dissimilarity measures provides a way of assessing their respective discriminant power for the data at hand. Also, using, for instance, the multichannel approach considered by Pollock (2007), the proposed methods could be applied on parallel sequences such as those describing, for example, linked lives or joint occupational and cohabitational trajectories. The observed differences between groups could then result from any of the channel, or from the combination of channels.[10] Even more generally, the discrepancy analysis is not limited to sequence data. If we except the graphical rendering of the results, they apply to any objects that can be characterized by a pairwise dissimilarity matrix.

Finally, we would like to remind our readers that all the proposed tools have been implemented in the TraMineR library for the R statistical environment. They are thus readily and freely accessible to any interested reader as illustrated in Section 9.

Acknowledgments:

We gratefully thank Nevena Zhelyazkova for her careful reading of the manuscript and the anonymous reviewers for their constructive comments and suggestions.

Funding

This research was supported by the Swiss National Science Foundation (Grant SNSF 100015-122230).

[10] It should be noted, however, that such a multichannel analysis is not intended for studying the link between the channels.

A Proofs

In this appendix, we present the mathematical developments underlying the results presented in the article. The presentation is largely inspired from Späth (1975) and Batagelj (1988).

We begin with a proof of Equation (1) that expresses the sum of squares in terms of pairwise Euclidean distances. Here we demonstrate it for the more general multivariate case. Let $\mathbf{y}_i$ be the data vector for case i, w_i its associated weight, $W = \sum_{i=1}^{n} w_i$ the sum of the weights, and $\bar{\mathbf{y}} = \frac{1}{W}\sum_{i=1}^{n} w_i \mathbf{y}_i$ the vector of weighted averages. Letting $\|\mathbf{y}\|^2$ denote the squared length of the vector, that is $\mathbf{y}'\mathbf{y} = \sum_i y_i^2$, the multivariate result that we want to establish is:

Theorem 1 ***Sum of squares in terms of pairwise distances***

$$SS \quad = \quad \sum_{i=1}^{n} w_i \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2 = \frac{1}{W}\sum_{i=1}^{n}\sum_{j=i+1}^{n} w_i w_j \|\mathbf{y}_i - \mathbf{y}_j\|^2 \tag{16}$$

Proof. We first show that the sum of squared distances to a point $\mathbf{x}$ is

$$\sum_{i=1}^{n} w_i \|\mathbf{y}_i - \mathbf{x}\|^2 \quad = \quad \sum_{i=1}^{n} w_i \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2 + W\|\bar{\mathbf{y}} - \mathbf{x}\|^2 \tag{17}$$

Since $\mathbf{y}_i - \mathbf{x} = (\mathbf{y}_i - \bar{\mathbf{y}}) + (\bar{\mathbf{y}} - \mathbf{x})$, its squared length is

$$\|\mathbf{y}_i - \mathbf{x}\|^2 \quad = \quad \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2 + 2(\bar{\mathbf{y}} - \mathbf{x})'(\mathbf{y}_i - \bar{\mathbf{y}}) + \|\bar{\mathbf{y}} - \mathbf{x}\|^2$$

Weighting with w_i and summing over i, we get

$$\sum_{i=1}^{n} w_i \|\mathbf{y}_i - \mathbf{x}\|^2 \quad = \quad 2(\bar{\mathbf{y}} - \mathbf{x})' \sum_{i=1}^{n} w_i (\mathbf{y}_i - \bar{\mathbf{y}}) + \sum_{i=1}^{n} w_i \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2 + W\|\bar{\mathbf{y}} - \mathbf{x}\|^2 \tag{18}$$

Since, $\sum_{i=1}^{n} w_i(\mathbf{y}_i - \bar{\mathbf{y}}) = \mathbf{0}$, the middle term on the right-hand side vanishes, which yields Equation (17). Setting $\mathbf{x} = \mathbf{y}_j$, multiplying by w_j and summing over j results in

$$\begin{aligned} \sum_{j=1}^{n}\sum_{i=1}^{n} w_i w_j \|\mathbf{y}_i - \mathbf{y}_j\|^2 \quad &= \quad \sum_{j=1}^{n} w_j \sum_{i=1}^{n} w_i \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2 + W\sum_{j=1}^{n} w_j \|\bar{\mathbf{y}} - \mathbf{y}_j\|^2 \\ &= \quad 2W \sum_{i=1}^{n} w_i \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2 \end{aligned} \tag{19}$$

The left-hand side can be written as $2\sum_{i=1}^{n}\sum_{j=i+1}^{n} w_i w_j \|\mathbf{y}_i - \mathbf{y}_j\|^2$. Then, dividing both sides by $2W$ we get Equation 16 of Theorem 1. □

Theorem 2 ***Contribution to the sum of squares.*** *It can be expressed as follows in terms of pairwise distances.*

$$\|\bar{\mathbf{y}} - \mathbf{x}\|^2 \quad = \quad \frac{1}{W}\Big(\sum_{i=1}^{n} w_i \|\mathbf{x} - \mathbf{y}_i\|^2 - SS\Big) \tag{20}$$

Proof. Extracting $\|\mathbf{x} - \bar{\mathbf{y}}\|^2$ from Equation (17), we get

$$\|\bar{\mathbf{y}} - \mathbf{x}\|^2 \quad = \quad \frac{1}{W}\Big(\sum_{i=1}^{n} w_i \|\mathbf{y}_i - \mathbf{x}\|^2 - \sum_{i=1}^{n} w_i \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2\Big) \tag{21}$$

The second term in the parenthesis is just SS, which proves the Theorem. □

Replacing $\mathbf{x}$ with $\mathbf{y}_j$, we can see that $\|\bar{\mathbf{y}} - \mathbf{y}_j\|^2$ is the contribution of $\mathbf{y}_j$ to SS by multiplying Equation (20) by w_j and summing over j. As a result, we obtain $\sum_j w_j \|\bar{\mathbf{y}} - \mathbf{y}_j\|^2 = 2SS - SS = SS$, hence the sum of squares. What makes the formula interesting is that it expresses the contribution in terms of pairwise distances. Formula (4 on page 5) is just Theorem 2 with dissimilarities d^{ν} substituted in place of the squared Euclidean distances $\|.\|^2$.

We now prove the following result about the non-negativity of the contribution.

Theorem 3 ***Non-negativity of the contribution to the sum of squares (sufficient condition).*** *Let d be a dissimilarity measure and assume the generalized sum of squares SS is calculated with d^ν. Then, the contribution of* $\mathbf{x}$ *to the sum of squares SS*

$$d^\nu_{x\tilde{g}} \quad = \quad \frac{1}{W}\Big(\sum_{i=1}^{n} w_i d^\nu_{xi} - SS\Big) \tag{22}$$

is non-negative when d^ν respects the triangle inequality.

Proof. Replacing SS by its expression in terms of pairwise dissimilarities (Theorem 1), the contribution (i.e., Equation (22)), can be re-written as

$$d^\nu_{x\tilde{g}} \quad = \quad \frac{1}{2W^2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j\big(2d^\nu_{ix} - d^\nu_{ij}\big) \tag{23}$$

In this form, it appears that the smallest value of the contribution $d^\nu_{x\tilde{g}}$ is obtained when all d^ν_{ij}'s take their maximal possible value. Under the triangle inequality, d^ν_{ij} cannot exceed $d^\nu_{xi} + d^\nu_{xj}$. Hence, $d^\nu_{x\tilde{g}}$ reaches its minimum when $d^\nu_{ij} = d^\nu_{xi} + d^\nu_{xj}$ for all i and j. This minimum is zero, which implies $d^\nu_{x\tilde{g}} \geq 0$. □

B Should dissimilarities be squared?

In this appendix, we discuss the choice of the ν exponent in Equation (2). Should we square the dissimilarities when computing the generalized sum of squares, or is it preferable to substitute the squared Euclidean distances with the dissimilarities themselves?

If the chosen dissimilarity between sequences can be represented univocally as a distance in an associated Euclidean coordinate space, we would have to set $\nu = 2$ to get a generalized SS equal to the corresponding sum of squares in that space (Gower 1982). While dissimilarities between strings of characters that can be expressed as kernels (Lodhi, Saunders, Shawe-Taylor, Cristianini, and Watkins 2002) have this property, most cost-minimizing distances such as OM cannot be expressed as Euclidean distances.

There are several arguments in favor of setting $\nu = 1$. According to Mielke and Berry (1983), this solution leads to a strongest congruence between analysis and data space. Moreover, it should produce more robust results when the corresponding points in the coordinate space are not normally distributed. From our point of view, the strongest argument to set $\nu = 1$ is related to the triangle inequality. Indeed, when the dissimilarity d respects the triangle inequality, $\sqrt{d}$ respects it too, while generally d^2 does not. Since the triangle inequality of d^ν ensures that SS cannot be greater than the sum of distances $\sum_i w_i d^\nu_{xi}$ to any arbitrary chosen object x, we would then be sure that SS does not exceed the sum $\sum_i w_i d^\nu_{xi}$ with $\nu = 1$, while it would not be the case with $\nu = 2$. The same argument can be formalized differently in terms of the contribution (4) to the sum of squares SS. The non-negativity of this contribution automatically results when d^ν_{ij} satisfies the triangle inequality (see Appendix A), while negative contributions to the discrepancy can occur when the triangle inequality does not hold. Hence, $\nu = 1$ ensures non-negative contributions when d satisfies the triangle inequality.

A negative value of the dissimilarity $d^\nu_{x\tilde{g}}$ between x and the center of gravity $\tilde{g}$ means that accounting for the object x reduces the sum of squares. This can be the case when two objects, say y and z, become closer when we can pass through x (i.e., when $d_{yz} > d_{yx} + d_{xz}$). Such situations are common in social network analysis. Consider, for instance, a network between x, y and z where the dissimilarity is equal to 1 for two people that meet often and is equal to 10 when they never meet. The dissimilarity $d_{x\tilde{g}}$ would then be negative if x often meets both y and z while y never meets z. From a social network perspective, we would say that x plays a cohesive role in the network. Although a negative contribution to the discrepancy is relevant in such settings, it is most often not the case. Hence, the results should be interpreted with caution when d^ν_{ij} does not respect the triangle inequality, which may occur with $\nu = 2$ as noted above. In particular, in such situations one should be ready to accept and give meaning to negative contributions to the discrepancy.

To summarize, we suggest defining SS with $\nu = 1$, except when we can express the dissimilarity measure as an Euclidean distance, in which case $\nu = 2$ is best suited.

C About the number of permutations in permutation tests

It is generally admitted that 1,000 permutations are sufficient to assess a result at the 5% level, while 5,000 are necessary at the 1% level. Here, we present some figures to support this claim.

Let p be the true p-value of F_{obs} and $\hat{p}$ be the proportion of F's smaller than F_{obs} among R randomizations. Table 6 shows how the probability $P(\hat{p} < 1.2p \mid p)$ that the empirical p-value does not exceed the true p-value by more than 20% evolves with R for $p = .05$ and $p = .01$. In the same table, we see that 1,000 randomizations ensure that this probability will be greater than 90% for $p = .05$, and $R = 5,000$ ensures this confidence for $p = .01$. The table shows also 95% inconclusive intervals—that is, the interval that should contain 95% of the permutation p-values when the true p-value is p. This interval is equal to $p \pm 1.96\sqrt{p(1-p)/R}$ (Manly 2007). Obtaining a permutation p-value in this interval does not permit one to conclude since such a p-values would then not be significantly different from p.

Table 6. Probability $P(\hat{p} < 1.2p \mid p)$ of Not Exceeding p by More Than 20% and 95% Inconclusive Interval When the True Value is p for a Selection of R Values

	$p = .05$			$p = .01$		
R	$P(\hat{p} < 1.2p \mid p)$	Inconclusive interval		$P(\hat{p} < 1.2p \mid p)$	Inconclusive interval	
100	0.677	0.007	0.093	0.580	0	0.030
200	0.742	0.020	0.080	0.612	0	0.024
500	0.848	0.031	0.069	0.673	0.001	0.019
1,000	0.927	0.036	0.064	0.737	0.004	0.016
1,300	0.951	0.038	0.062	0.766	0.005	0.015
5,000	0.999	0.044	0.056	0.922	0.007	0.013
7,000	1.000	0.045	0.055	0.954	0.008	0.012
10,000	1.000	0.046	0.054	0.978	0.008	0.012
50,000	1.000	0.048	0.052	1.000	0.009	0.011
100,000	1.000	0.049	0.051	1.000	0.009	0.011

References

Abbott, Andrew. 1990. "A Primer on Sequence Methods." *Organization Science* 1:375–392.

Abbott, Andrew and John Forrest. 1986. "Optimal Matching Methods for Historical Sequences." *Journal of Interdisciplinary History* 16:471–494.

Abbott, Andrew and Alexandra Hrycak. 1990. "Measuring Resemblance in Sequence Data: An Optimal Matching Analaysis of Musician's Carrers." *American Journal of Sociolgy* 96:144–185.

Anderson, Marti Jane. 2001. "A new method for non-parametric multivariate analysis of variance." *Austral Ecology* 26:32–46.

Anderson, Marti Jane. 2006. "Distance-Based Tests for Homogeneity of Multivariate Dispersions." *Biometrics* 62:245–253.

Bartlett, Maurice Stevenson. 1937. "Properties of Sufficiency and Statistical Tests." *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 160:268–282.

Batagelj, Vladimir. 1988. "Generalized Ward and related clustering problems." In *Classification and related methods of data analysis*, edited by Hans H. Bock, pp. 67–74. Amsterdam: North-Holland.

Billari, Francesco Candeloro. 2001a. "The Analysis of Early Life Courses: Complex Description of the Transition to Adulthood." *Journal of Population Research* 18:119–142.

Billari, Francesco Candeloro. 2001b. "A Log-Logistic Regression Model for a Transition Rate with a Starting Threshold." *Population Studies* 55:15–24.

Billari, Francesco Candeloro. 2005. "Life Course Analysis: Two (Complementary) Cultures? Some Reflections With Examples From The Analysis Of Transition To Adulthood." In *Towards an Interdisciplinary Perspective on the Life Course*, edited by René Levy, Paolo Ghisletta, Jean-Marie Le Goff, Dario Spini, and Eric Widmer, Advances in Life Course Research, Vol. 10, pp. 267–288. Amsterdam: Elsevier.

Blossfeld, Hans-Peter and Götz Rohwer. 2002. *Techniques of Event History Modeling, New Approaches to Causal Analysis*. Mahwah NJ: Lawrence Erlbaum, 2nd edition.

Breiman, Leo, Jerome H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification And Regression Trees*. New York: Chapman and Hall.

Brown, Morton B. and Alan B. Forsythe. 1974a. "Robust Tests for the Equality of Variances." *Journal of the American Statistical Association* 69:364–367.

Brown, Morton B. and Alan B. Forsythe. 1974b. "The Small Sample Behavior of Some Statistics Which Test the Equality of Several Means." *Technometrics* 16:129–132.

Cuadras, Carles M. 2008. "Distance-Based Association and Multi-Sample Tests for General Multivariate Data." In *Advances in Mathematical and Statistical Modeling*, edited by Barry C. Arnold, N. Balakrishnan, Jose-Maria Sarabia, and Roberto Minguez, Statistics for Industry and Technology. Birkhäuser Boston.

Delicado, Pedro. 2007. "Functional k-sample problem when data are density functions." *Computational Statistics* 22:391–410.

Dijkstra, Will and Toon Taris. 1995. "Measuring the Agreement between Sequences." *Sociological Methods and Research* 24:214–231.

Elder, Glen H. 1999. *Children of the Great Depression*. Westview Press.

Elzinga, Cees H. 2003. "Sequence Similarity: A Non-Aligning Technique." *Sociological Methods and Research* 31:214–231.

Elzinga, Cees H. 2007. "Sequence Analysis: Metric Representations of Categorical Time Series." Manuscript, Dept of Social Science Research Methods, Vrije Universiteit, Amsterdam.

Elzinga, Cees H. 2010. "Complexity of categorical time series." *Sociological Methods and Research* 38:463–481.

Elzinga, Cees H. and Aart C. Liefbroer. 2007. "De-standardization of Family-Life Trajectories of Young Adults: A Cross-National Comparison Using Sequence Analysis." *European Journal of Population* 23:225–250.

Gabadinho, Alexis, Gilbert Ritschard, Nicolas S Müller, and Matthias Studer. 2011a. "Analyzing and Visualizing State Sequences in R with TraMineR." *Journal of Statistical Software* 40:1–37.

Gabadinho, Alexis, Gilbert Ritschard, Matthias Studer, and Nicolas S. Müller. 2009. "Mining Sequence Data in R with the TraMineR package: A User's Guide." Technical report, Department of Econometrics and Laboratory of Demography, University of Geneva, Geneva.

Gabadinho, Alexis, Gilbert Ritschard, Matthias Studer, and Nicolas S. Müller. 2010. "Indice de complexité pour le tri et la comparaison de séquences catégorielles." *Revue des nouvelles technologies de l'information RNTI* E-19:61–66.

Gabadinho, Alexis, Gilbert Ritschard, Matthias Studer, and Nicolas S. Müller. 2011b. "Extracting and

Rendering Representative Sequences." In *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, edited by Ana Fred, Jan L. G. Dietz, Kecheng Liu, and Joaquim Filipe, volume 128 of *Communications in Computer and Information Science (CCIS)*, pp. 94–106. Springer-Verlag.

Gansner, Emden R. and Stephen C. North. 1999. "An Open Graph Visualization System and Its Applications to software engineering." *Software - Practice and Experience* 30:1203–1233.

Geurts, Pierre, Louis Wehenkel, and Florence d'Alché Buc. 2006. "Kernelizing the output of tree-based methods." In *ICML*, edited by William W. Cohen and Andrew Moore, volume 148 of *ACM International Conference Proceeding Series*, pp. 345–352. ACM.

Gower, John Clifford. 1966. "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis." *Biometrika* 53:325–338.

Gower, John Clifford. 1982. "Euclidean Distance Geometry." *Mathematical Scientist* 7:1–14.

Gower, John Clifford and Wojtek J. Krzanowski. 1999. "Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 48:505–519.

Jobson, J. D. 1991. *Applied Multivariate Data Analysis*, volume I: Regression and Experimental design. New York: Springer-Verlag.

Lesnard, Laurent. 2010. "Setting Cost in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns." *Sociological Methods and Research* 38:389–419.

Lodhi, Huma, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. "Text classification using string kernels." *The Journal of Machine Learning Research* 2:419–444.

Manly, Bryan F. J. 2007. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman & Hall, third edition edition.

McArdle, Brian H. and Marti J. Anderson. 2001. "Fitting Multivariate Models to Community Data: A Comment on Distance-Based Redundancy Analysis." *Ecology* 82:290–297.

McVicar, Duncan and Michael Anyadike-Danes. 2002. "Predicting successful and unsuccessful transitions from school to work using sequence methods." *Journal of the Royal Statistical Society A* 165:317–334.

Mielke, Paul W. and Kenneth J. Berry. 1983. "Asymptotic clarifications, generalizations, and concerns regarding an extended class of matched pairs tests based on powers of ranks." *Psychometrika* 48:483–485.

Mielke, Paul W. and Kenneth J. Berry. 2007. *Permutation Methods: A Distance Function Approach*. New York: Springer, 2nd edition edition.

Morgan, J. N. and J. A. Sonquist. 1963. "Problems in the Analysis of Survey Data, and a Proposal." *Journal of the American Statistical Association* 58:415–434.

Piccarreta, Raffaella. 2010. "Binary trees for dissimilarity data." *Computational Statistics and Data Analysis* 54:1516–1524.

Piccarreta, Raffaella and Francesco Candeloro Billari. 2007. "Clustering work and family trajectories by using a divisive algorithm." *Journal of the Royal Statistical Society A* 170:1061–1078.

Pollock, Gary. 2007. "Holistic trajectories: A study of combined employment, housing and family careers by using multiple-sequence analysis." *Journal of the Royal Statistical Society A* 170:167–183.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reiss, Philip T, M. Henry H Stevens, Zarrar Shehzad, Eva Petkova, and Michael P Milham. 2009. "On Distance-Based Permutation Tests for Between-Group Comparisons." *Biometrics* 66:636–43.

Scherer, Stefani. 2001. "Early Career Patterns: A Comparison of Great Britain and West Germany." *European Sociological Review* 17:119–144.

Shaw, Ruth G. and Thomas Mitchell-Olds. 1993. "Anova for Unbalanced Data: An Overview." *Ecology* 74:1638–1645.

Späth, Helmuth. 1975. *Cluster analyse algorithmen*. München: R. Oldenbourg Verlag.

Studer, Matthias, Gilbert Ritschard, Alexis Gabadinho, and Nicolas S. Müller. 2009. "Analyse de dissimilarités par arbre d'induction." *Revue des nouvelles technologies de l'information RNTI* E-15:7–18.

Studer, Matthias, Gilbert Ritschard, Alexis Gabadinho, and Nicolas S. Müller. 2010. "Discrepancy analysis of complex objects using dissimilarities." In *Advances in Knowledge Discovery and Management*, edited by Fabrice Guillet, Gilbert Ritschard, Djamel A. Zighed, and Henri Briand, volume 292 of *Studies in Computational Intelligence*, pp. 3–19. Berlin: Springer.

Widmer, Eric and Gilbert Ritschard. 2009. "The De-Standardization of the Life Course: Are Men and Women Equal?" *Advances in Life Course Research* 14:28–39.

Wu, Lawrence L. 2000. "Some Comments on 'Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect'." *Sociological Methods Research* 29:41–64.

Yujian, Li and Liu Bo. 2007. "A Normalized Levenshtein Distance Metric." *IEEE Transactions On Pattern Analysis And Machine Intelligence* 29:1091–1095.

Zapala, Matthew A. and Nicholas J. Schork. 2006. "Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables." *Proceedings of the National Academy of Sciences of the United States of America* 103:19430–19435.

Matthias Studer is a PhD student in Socioeconomics and a research assistant at the Institute for Demographic and Life Course Studies at the University of Geneva. He holds a Master degree in economics and a Master of Advanced Studies in sociology. He is currently working on gender and social inequalities at the beginning of academic careers in Switzerland. His research interests include data mining of longitudinal data such as state and event sequences, dissimilarity analysis and survival trees.

Gilbert Ritschard is a full professor of statistics for the social sciences at the University of Geneva. His current research interests are in life course analysis and the application of exploratory data mining methods in social sciences. Beside many contributions to chapter books, He published recently in *Advanced Life Course Research*, *Studies in Family Planning* and in the *International Journal of Data Mining, Modelling and Management*, and co-edited a book on *Advances in Knowledge Discovery and Management*. He leads currently a Swiss NSF research project on methods for mining event histories.

Alexis Gabadinho is a scientific collaborator at the Institute for Demographic and Life Course Studies at the University of Geneva. He holds a post-graduate diploma in demography. His current research interests are the application of data mining methods in social sciences and the development of methods for categorical state sequences analysis, in particular measures of sequence complexity and methods for summarizing sets of sequences.

Nicolas S. Müller is currently a PhD student in Sociology and a research assistant at the Institute for Demographic and Life Course Studies at the University of Geneva. He holds a MA in sociology and a MSc in Information Systems. His PhD subject is about the links between life trajectories, socio-economic factors and health outcomes. He is interested in the application of data mining methods in social sciences, and especially sequence mining and association rules methods.

The authors are the developers of the TraMineR toolbox for analyzing sequence data in R.