Article

What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures

STUDER, Matthias, RITSCHARD, Gilbert

Abstract

This is a comparative study of the multiple ways of measuring dissimilarities between state sequences. The originality of the study is the focus put on the differences between sequences that are sociologically important when studying life courses such as family life trajectories or professional careers. These differences essentially concern the sequencing (the order in which successive states appear), the timing and the duration of the spells in successive states. The study examines the sensitivity of the measures to these three aspects analytically and empirically by means of simulations. Even if some distance measures underperform, the study shows that there is no universally optimal distance index, and that the choice of a measure depends on which aspect we want to focus on. From the review and simulation results, the paper derives guidelines to help the end user to choose the right dissimilarity measure for her or his research objectives. This study also introduces novel ways of measuring dissimilarities that overcome some flaws in existing measures.

Reference

STUDER, Matthias, RITSCHARD, Gilbert. What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society. Series A. Statistics in society*, 2016, vol. 179, no. 2, p. 481–511

DOI : 10.1111/rssa.12125

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2016)

What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures

Matthias Studer and Gilbert Ritschard

University of Geneva, Switzerland

[Received March 2014. Final revision March 2015]

Summary. This is a comparative study of the multiple ways of measuring dissimilarities between state sequences. The originality of the study is the focus put on the differences between sequences that are sociologically important when studying life courses such as family life trajectories or professional careers. These differences essentially concern the sequencing (the order in which successive states appear), the timing and the duration of the spells in successive states. The study examines the sensitivity of the measures to these three aspects analytically and empirically by means of simulations. Even if some distance measures underperform, the study shows that there is no universally optimal distance index, and that the choice of a measure depends on which aspect we want to focus on. From the review and simulation results, the paper derives guidelines to help the end user to choose the right dissimilarity measure for her or his research objectives. This study also introduces novel ways of measuring dissimilarities that overcome some flaws in existing measures.

Keywords: Dissimilarity; Distance; Duration; Optimal matching; Sequencing; Spells; State sequences; Timing

1. Introduction

Abbott (1983) stressed the relevance of sequence methods to the social sciences and founded theoretically the use of sequence analysis on narrative positivism (Abbott, 1992). Since then, sequence analysis has become popular, and particularly so-called optimal matching (OM) analysis (Abbott and Forrest, 1986; Abbott and Hrycak, 1990). Sequence analysis is now a key method used to study the spans of life trajectories and careers (e.g. Bras *et al*. (2010), Widmer and Ritschard (2009) and Schumacher *et al*. (2012)). The strength of the sequence approach is the holistic view that it provides by dealing with whole trajectories. This allows us to determine trajectory patterns that account for all states of interest experienced during the period considered. In contrast, survival or event history analyses focus on the hazard of—or time to—a specific event, and do not give an overall view of how the trajectories are organized.

An OM analysis measures pairwise dissimilarities between sequences and then identifies 'types' of pattern by clustering the sequences based on these dissimilarities. Beneath clustering analyses, other dissimilarity-based methods have also proven useful when investigating sequence data. For instance, Abbott (1983) mentioned multi-dimensional scaling, Massoni *et al*. (2009) used self-organizing maps, Studer *et al*. (2011) showed how to run analysis-of-variance like

Address for correspondence: Matthias Studer, Institute for Demographic and Life Course Studies, University of Geneva, Boulevard du Pont d'Arve 40, CH-1211 Geneva 4, Switzerland.
E-mail: Matthias.Studer@unige.ch

analyses and to grow regression trees on sequence data, and Gabadinho and Ritschard (2013) searched for non-redundant typical patterns with the densest neighbourhoods.

Despite often being referred to as 'OM analysis', so named by Abbott and Forrest (1986) after the edit distance they used, a dissimilarity-based analysis is in no way restricted to OM distances. The methods also work with other measures of dissimilarity, and, as we shall see, many different distances have been proposed. For example, there are χ^2-distances that have been adapted for sequence data, which essentially measure differences in state distributions (e.g. Deville and Saporta (1983) and Grelet (2002)). There are also distances based on counts of common attributes, e.g. matching states (Hamming, 1950; Bergroth *et al.*, 2000) or matching subsequences (Elzinga and Studer, 2015), and multiple variants of editing dissimilarity measures, such as OM, that evaluate differences according to the cost of 'editing' one sequence into the other (e.g. Levenshtein (1966), Hollister (2009), Halpin (2010), Lesnard (2010) and Biemann (2011)).

Measuring the dissimilarity between sequences (i.e. a pairwise comparison of the sequences) is the common and crucial starting point for all dissimilarity-based sequence methods. Therefore, in choosing a dissimilarity measure, it is important that we understand what we want to account for before quantitatively evaluating the difference between two sequences. This study contributes to this understanding by identifying the aspects (e.g. constituent states, sequencing, timing and duration) in which sequences may differ, and studying how various dissimilarity measures account for these aspects. This study of dissimilarity measures comprises an organized descriptive review and is original in that it focuses on those aspects of sequence differences that matter in the social sciences. In addition, we conduct a simulation study to examine how these measures behave with respect to those aspects.

Alongside this review, we propose two new distance measures and three original strategies to set OM costs. The first new distance measure is an edit measure—OM between sequences of spells—that consistently accounts for differences in the time that is spent in the distinct successive states (DSSs). The second is a reformulation of the OM of transition sequences that was introduced by Biemann (2011). The variant proposed drastically reduces the number of parameters. With regard to setting the OM costs, we first propose an original solution to set data-driven insertion and deletion, indel, costs based on state frequencies. Second, we suggest the use of the Gower distance to determine the substitution costs for a mix of quantitative and qualitative state attributes. Finally, we propose to define substitution costs as a χ^2-distance that stresses the similarity between states sharing the same future.

The remainder of this paper is organized as follows. We first set the framework by specifying the kinds of sequences that we consider, and the different aspects that we may want the dissimilarity measures to reflect. We then present the dissimilarity measures reviewed and their theoretical properties. In the following section, we examine the behaviour of the measures by using artificially generated data, and we empirically study how the measures are related to each other. Finally, we conclude the paper by providing guidelines on how to select an appropriate measure.

All measures that are presented in this paper are available in the latest version of the `TraMineR` R library. See the `TraMineR` Web site (`http://traminer.unige.ch`) for explanations on how to compute these measures.

2. Sequences and distances

2.1. Definitions and notation

We consider categorical sequences, defined as an ordered list of successive elements chosen from a finite alphabet, Σ. For sequences describing life trajectories, the elements in the sequences are

usually in chronological order. In addition, in discrete time state sequences, the position in the sequence conveys time information so that the difference between two positions defines a duration. For example, assuming positions correspond to ages in years, knowing that an individual is in state 'full-time work' from positions 20 to 29, we can conclude that the individual worked full time for 10 years.

The natural way to encode a state sequence is to list the successive elements. For example, the trajectory of someone working 'full time', F, for 2 years and then 'part time', P, for 3 years is represented as F–F–P–P–P. We can also encode the sequence in a more compact way as F^2–P^3. In other words, we simply list the DSSs and add a duration stamp—in this case, as the superscript—indicating the number of successive positions in that state (i.e. the length of the spell in the state). Apart from being compact, the latter form also facilitates comparing the sequencing and duration of spells in the same state. In the remainder of this study, we shall use the term *spell* to refer to the whole spell spent in the same state.

2.2. Differences between sequences

State sequences are complex objects that provide many different pieces of information, such as total and consecutive time spent in each state, the timing of states and the state order. Kruskal (1983), page 207, distinguished four different ways—or transformation operations—in which sequences may differ: substitutions, indels, compressions and expansions, and transpositions or swaps. These transformation-based distinctions make sense in fields such as biology, computer science and speech research, and motivated the basic operations that are considered in edit distances. In the social sciences, the life trajectory of one individual can hardly be considered to be the result of a transformation of the trajectory of another person. Therefore, our interest when comparing sequences is not in the transformation of one sequence to another, but in how the sequences differ in socially meaningful aspects. In line with the distinctions that were made by Settersten and Mayer (1997) and Billari *et al*. (2006), we identify the following important aspects:

(a) *experienced states*—the distinct elements of the alphabet present in the sequence;
(b) *distribution*—the within-sequence state distribution (total time);
(c) *timing*—the age or date at which each state appears;
(d) *duration*—the spell lengths in the distinct successive states;
(e) *sequencing*—the order of the distinct successive states.

The first basic aspect of interest when comparing the trajectories of two individuals is the *list of distinct states that each experiences*. This is what Dijkstra and Taris (1995) and Elzinga (2003) implicitly referred to when stating that two sequences with no common state are maximally dissimilar. (Note that this claim would not hold if some states can be considered to be more similar than others.) The notion of experienced states is also related to the quantum that was defined by Billari *et al*. (2006) as the count of experienced events. In addition to the list of experienced states, we may want to examine the total time spent in each distinct state. This tells us the *distribution* of the states within each sequence. Knowing the distribution is useful, for example, when studying the effect of total exposure times. For instance, we may want to examine the effect that the total amount of time spent unemployed has on a person's health status at retirement. However, differences in the presence or absence of states, or in the distribution of the states within the sequences, do not account for how the states occur along the longitudinal axis. Therefore, in a sequence analysis, these differences should be used in conjunction with other dimensions.

The *timing* of the states (i.e. the age—or date—at which we are in a given state) or the time that events occur, such as the start of a spell (Settersten and Mayer, 1997) in a given state, is a sociologically important aspect. For instance, life course literature often stresses the role of age norms in the construction of life trajectories (Widmer *et al*., 2003). Moreover, the social reality that is reflected by a state often depends on its position in the trajectory. For example, Rousset *et al*. (2012) observed that the effect of unstable employment on the professional integration trajectory increases with age. In addition, in his study on the way that couples use time, Lesnard (2010) claimed that differences between 'no partner working' and 'only one partner working' reflect a very different reality when observed during the day or night. Then, *spell duration*, the consecutive time that is spent in the same state, is another way to account for time (see Settersten and Mayer (1997), who even considered the more general concept of *spacing* to refer to the time between any two events or transitions). Instead of the precise timing, spell duration refers to the time that elapses between the start and the end of a significant spell. The spell durations, such as the time lived alone before marrying, or the duration of a jobless episode, are important aspects within people's life courses. Spell duration is different from the information that is provided by the state distribution in that it gives the consecutive exposure time, rather than the total, but not necessarily consecutive exposure time. Unlike the state distribution, the spell duration allows us, for example, to distinguish between long-term unemployment and multiple short-term unemployment episodes.

Finally, *sequencing*, the order in which states (or events) are experienced, is another socially sound dimension. The role of sequencing norms in the construction of life trajectories is at least as important as the role of age norms and has been emphasized by, for example, Hogan (1978). For example, experiencing childbirth before or after marriage reflects different ways of life. Abbott (1990) identified sequencing as the key concept in sequence analysis, and Billari *et al*. (2006) emphasized its importance in conjunction with timing and quantum for demographic life course analysis.

The five aforementioned aspects are not independent of each other. For example, by changing the sequencing, we also change the timing. Similarly, changing consecutive times spent in states implies changes in the within-sequence distribution, and possibly in the sequencing as well. Likewise, modifying the within-sequence distribution by changing the distinct present states affects the sequencing and duration. From the reverse point of view, two sequences that are similar with regard to one aspect may be quite different in terms of another.

In fact, we do not need all five aspects to characterize a sequence entirely. Specifying the sequencing (the DSS), the duration of the DSS and at least one time (e.g. the start time of the sequence) automatically determines the experienced states, their distribution and their time of occurrence. Likewise, the sequencing and the start time of the successive spells completely define the sequence. Therefore, from here on, we essentially focus on sequencing, duration and timing aspects.

In practice, we compare two sequences by using a measure of dissimilarity to quantify the level of mismatch between the sequences. The next section reviews existing dissimilarity measures, while stressing their properties and their sensitivity to timing, duration and sequencing. Among the properties, we shall in particular pay attention to the fulfilment of the mathematical conditions of a metric distance. These are required for most applications and especially for sample-based studies. Discrepancy analyses and many clustering algorithms (such as Ward) require metrics. For instance, the triangle inequality ensures coherence between computed dissimilarities. Without the triangle inequality, the actual dissimilarity between x and y could be smaller than the measured dissimilarity $d(x, y)$ because of a third sequence z. In this case, the actual dissimilarity would depend on the other sequences in the data set (see, for example,

Elzinga and Studer (2015). Among other cases, this is problematic in sample-based studies, where the actual distance depends on whether the z-sequence was drawn or not.

3. Overview of dissimilarity measures

Table 1 lists the dissimilarities reviewed. The first column gives short names, which will be used later when presenting the results of our empirical evaluations. The dissimilarity measures can be classified into three broad classes, as shown in the next three columns:

(a) distances between distributions, 'Dis';
(b) measures based on the count of common attributes between sequences, 'Att';
(c) edit distances, which measure the cost of the operations that are necessary to transform one sequence into the other, 'Edt'.

The next five columns indicate the properties of the measures. 'Metric' denotes measures fulfilling the mathematical conditions of distances. Then, 'Eucl' denotes Euclidean distances, 'T.warp' denotes measures allowing for a time warp when comparing sequences, 'S.dep' denotes state-dependent measures (i.e. measures that allow for differences between states that can vary) and 'Ctxt' denotes measures that consider the context of the states.

These properties may help to narrow the set of potentially useful distances. For example, we may want to discard OM with so-called optimized costs, OM(opt), because of the possible negative values that it can generate. We may also want to discard non-metric measures such as OM with transition-based costs, OM(trate), localized OM, OMloc, and dynamic Hamming costs, DHD, with costs derived from transition rates, because of unexpected behaviour that may result from possible violations of the triangle inequality. In addition, Euclidean distances may be preferred if we plan to use multi-dimensional scaling. For non-Euclidean distances, multi-dimensional scaling produces complex co-ordinates that are associated with negative eigen-values. These (usually ignored) complex co-ordinates reflect the distortion that is incurred by embedding sequences in a Euclidean vector space. Therefore, it may be worth studying these in further detail (Laub and Müller, 2004).

The last columns in Table 1 show the available tuning parameters. These parameters are explained in the following subsections, where each measure is briefly described. Here, 'Subst' represents the possibility of accounting for state-dependent substitution—or proximity—costs, with 'User' meaning that the costs are set by the user, 'Data' that they are data driven and 'Features' that they are based on state features. The 'Indels' column indicates whether there is a single state-independent indel cost, 'Single', whether state-dependent user-defined indel costs are allowed, 'Multiple', or whether the indel costs are automatically set by the measure itself, 'Auto'.

Considering state proximities or substitution costs is of special interest when some states should obviously be considered closer than others. Such distinctions occur, for instance, when the states are ordinal, such as education level, or result when some states share a higher number of common attributes than others. The possibility of considering state-dependent substitution costs is also of interest in the multichannel case. The method that was adopted by Pollock (2007) for measuring distances between multichannel sequences derives the multichannel costs from the costs that are available for each individual channel. In this case, the costs generated would at least vary with the number of channels concerned by the state mismatch. With the same fixed cost in each channel, for example, the cost will be lower for a difference in one channel only than for a simultaneous mismatch in two channels.

We now briefly describe each of the dissimilarity measures that are presented in Table 1. We start by addressing distances between within-sequence state distributions; then we consider

Table 1. Summary of dissimilarity measures between state sequences

Measure	*Type*			*Description*	*Properties*					*Parameters*		
	Dis	*Att*	*Edt*		*Metric*	*Eucl*	*T.warp*	*S.dep*	*Ctxt*	*Subst.*	*Indels*	*Others*
CHI2, EUCLID (Deville and Saporta, 1983)	×			Distance between per-period state distributions	×	×	×					Number of periods K
CHI2fut (Rousset *et al.*, 2012)	×			Positionwise state distances based on shared future	×	×			×			Time lag weighting function
NMS (Elzinga, 2003, 2005)		×		Based on number of matching subsequences	×	×	×		×			
SVRspell (Elzinga and Studer, 2015)		×		Based on number of matching spell subsequences with spell length weights	×	×	×	×	×	User		Subsequence length weight a; spell duration weight b
HAM (Hamming, 1950)		×	×	Number of mismatches	×	×†						
Generalized			×	Sum of mismatches with state-dependent weights	×‡	×†§		×		User		
DHD (Lesnard, 2010)			×	Sum of mismatches with positionwise state-dependent weights				×	×	Data		
OM (Abbott and Forrest, 1986)			×	Minimum cost for turning x into y by using theoretically defined costs	×‡		×	×		User	Multiple	
LCS, OM(1,2) or Levenshtein-II		×	×	Based on length of LCS or number of indels	×		×					
Feature (new)			×	Costs based on state features	×		×	×		Features	Single	State features
Future (new)			×	Costs based on similarity between conditional state distributions q periods ahead	×		×	×		Data	Single	Forward lag q
trate (Rohwer and Poetter, 2005)			×	Costs based on transition rates			×	×		Data	Single	Transition lag q
opt* (Gauthier *et al.*, 2009)			×	Costs adjusted to increase similarity between similar sequences	§§		×	×		Data	Single	Similarity rate
indels, indelslog (new)			×	State-dependent indels based on inverse or log-inverse state frequencies.	×		×	×			Auto	

(*continued*)

Table 1 (*continued*)

Measure	*Type*			*Description*	*Properties*					*Parameters*		
	Dis	*Att*	*Edt*		*Metric*	*Eucl*	*T.warp*	*S.dep*	*Ctxt*	*Subst.*	*Indels*	*Others*
OMloc (Holister, 2009)			×	Context-dependent indel costs			×	×	×	User	Auto	Expansion cost e; context g
OMslen (Halpin, 2010)			×	Costs weighted by spell length	×		×	×	×	User	Multiple*	Spell length weight h
OMspell (new)			×	OM between sequences of spells	×‡		×	×	×	User	Multiple*	Expansion cost e
OMstran (new)			×	OM between sequences of transitions	×‡		×	×	×	User	Multiple	Origin–transition trade-off w; transition indel cost function

†Squared Euclidean distance.
‡If costs fulfil the triangle inequality.
§If costs are squared Euclidean distances.
§§Can generate negative dissimilarities.
*Not available in `TraMineR`.

measures based on the count of common attributes. Lastly, we discuss OM and other related edit dissimilarities. In addition, we introduce two new distances measures to overcome problems that are identified in existing measures as well as three new strategies to define the costs in OM.

A more detailed and formalized review of the dissimilarities that are considered here can be found in Studer and Ritschard (2014). This working paper also provides a more thorough discussion of the sociological interpretation of the distances.

3.1. Distances between probability distributions

3.1.1. Distances between state distributions

One approach to measuring the dissimilarity between sequences, propounded by adepts of the French school of data analysis (Deville and Saporta, 1983; Grelet, 2002), focuses on the longitudinal state distribution within each sequence. In other words, the approach focuses on the time that is spent in each state within the sequences. The dissimilarity between sequences is measured by the distance between the distribution vectors by using either the Euclidean distance or the χ^2-distance. The former accounts for the absolute differences in the proportion of time spent in the states. The squared χ^2-distance weights the squared differences for each state by the inverse of the overall proportion of time spent in the state, which, for two identical differences, places more importance on a rare state than on a frequent state.

This first distribution-based measure is, by definition, sensitive to the time spent in the states. However, it is insensitive to the order and exact timing of the states. Following Deville and Saporta (1983), we can overcome this limitation by considering the distribution in K successive—possibly overlapping—periods. The distance is then equal to the sum of the χ^2-distances for each period. At the limit, when K is equal to the length of the sequences, the distance corresponds to a weighted count of mismatching states. The latter case will be very sensitive to non-matching timings and, as a result, gains some sensitivity to sequencing.

3.1.2. Distance based on conditional distributions of subsequent states

A related measure is defined as the sum of the position-dependent distances computed at successive positions. This measure was proposed by Rousset *et al*. (2012) to measure the dissimilarity between sequences describing professional integration trajectories. The aim of the measure is to stress the similarity of the sequences that are likely to lead to the same future. For example, two different educational trajectories will be considered similar if they are both likely to lead to the same stable professional position. Here, the distance between states at position t is itself defined as the χ^2-distance between the vectors of the (weighted and normalized) transition rates from the state observed at t to the states observed at the subsequent positions, $t+1, t+2, \ldots$. Each transition rate is weighted by a decreasing function of the time interval to give more importance to the near future than to the far future when evaluating the distance between the states at position t.

Since this distance is the sum of positionwise distances, it should be sensitive to non-matching timings and differences in sequencing. However, we can expect this sensitivity to be smoothed somewhat by the introduced link to the future.

The χ^2-distances are Euclidean, as is the sum of the Euclidean distances over positions. Therefore, all three distances are Euclidean and have all the desired mathematical properties. However, the distances that are defined as the sum of the positionwise distances between states apply only to pairs of sequences of the same length.

3.2. Distances based on counts of common attributes

3.2.1. Simple Hamming distance

Hamming (1950) proposed measuring the dissimilarity between two sequences by using the number of positions with non-matching states—the Hamming distance also corresponds to the Gower distance with equally weighted states and positions, as considered by Wilson (2006). Since the Hamming distance proceeds by a positionwise comparison, it applies only to pairs of sequences of the same length and is very sensitive to timing mismatches. The square root of the measure is Euclidean and, in its original formulation, is independent of the mismatching tokens.

3.2.2. Length of the longest common subsequence

The length of the longest common subsequence (LCS) corresponds to the number of elements in one sequence that can be uniquely matched with elements occurring in the same order in the other sequence (for example see Bergroth *et al*. (2000)). Letting $A(x, y)$ be the number of elements matching in this way, we obtain the LCS distance by computing $d_{\mathrm{LCS}} = A(x, x) + A(y, y) - 2A(x, y)$. Since the position in the other sequence with which an element is matched varies with the other sequence, d_{LCS} is not Euclidean. Moreover, since it is not based on positionwise matches, the LCS distance should not be too sensitive to timing. In this case, we can expect a stronger dependence on differences in the state distribution and sequencing, especially the order of the most frequent states and, to a lesser extent, to differences in the consecutive times spent in the distinct states.

3.2.3. Number of matching subsequences

Elzinga (2003, 2005) introduced a dissimilarity measure based on the number of matching subsequences, NMS. (A subsequence is obtained by deleting any number of states in a sequence (Elzinga *et al*., 2008).) The general idea of the measure is that, the more often a given ordering of tokens in one sequence is observed in the other sequence, the closer the two sequences are to each other.

Elzinga and Studer (2015) proposed a generalization of NMS called the *subsequence vector representation-based metric*, SVRspell. This distance is based on the matching subsequences between DSS sequences where the matching subsequences are weighted according to their length and the duration of the spells involved. Two parameters control the behaviour of the measure. The first parameter, $a \geqslant 0$, is an exponent for the subsequence length weights. The second parameter, $b \geqslant 0$, is an exponent for the spell durations. In addition to these weighting mechanisms, the subsequence vector representation SVR can account for state proximities.

NMS and SVRspell are Euclidean distances. They should be very sensitive to differences in sequencing and sensitive to differences in duration. Owing to the duration extension, the original version increases the number of embeddings of subsequences concerned. The second form does so by explicitly considering the duration of spells. In contrast, computing NMS between the DSS sequences, which is equivalent to SVRspell with $b = 0$, should be insensitive to differences in timing and duration.

3.3. Optimal matching

Since Andrew Abbott (Abbott and Forrest, 1986; Abbott and Hrycak, 1990) popularized OM analysis in the social sciences, OM has become the most common way of computing dissimilarities between sequences describing life trajectories. The method borrows from other fields

that use similar edit approaches (Kruskal, 1983), such as the Levenshtein distance (Levenshtein, 1966) in computer science and sequence alignment in bioinformatics.

3.3.1. Optimal matching principles and special cases

OM measures the dissimilarity between two sequences, x and y, as the minimum total cost of transforming one sequence, say x, into the other sequence y, by means of indels of tokens or substitutions between tokens. Each operation is assigned a cost, which may vary with the states involved.

The costs can be specified by using a single matrix, denoted as $\mathbf{\Gamma}$, where the indel costs are specified as a substitution with an additional 'null' or 'empty' state. The OM distance between the sequences is a metric if $\mathbf{\Gamma}$ defines a metric between the admissible states (Yujian and Bo, 2007). In other words, the costs should be symmetric, fulfil the triangle inequality and be 0 only for the substitution of an element with itself. If the triangle inequality is not satisfied, at least one substitution cost will not make sense, because there will be a path allowing the same substitution result at a lower cost. Moreover, existing algorithms, such as that of Needleman and Wunsch (1970), that are used to compute the OM distance all assume that the costs satisfy the metric properties. Therefore, they could return a solution that does not reflect the minimum cost if these properties are violated. As the solution to a minimization process, the OM distance cannot be expressed as a kernel and, therefore, is not Euclidean (Elzinga, 2007).

The parameterization of OM by using the costs of the elementary operations makes it a very flexible dissimilarity measure that can cope with many situations. It defines a range of distance measures between two extreme cases (Lesnard, 2010): the generalized Hamming distance and the Levenshtein II distance. The former is the weighted sum of positionwise mismatches between two sequences (i.e. OM without indels). Like the simple Hamming distance, it should be mainly sensitive to timing differences. The latter case is the number of indels that are needed to transform one sequence into the other (i.e. OM without substitutions). Note that, for a single-indel cost of 1, OM without substitutions (Levenshtein II) is equivalent to OM with substitution costs of 2 or more, and to the LCS distance. The Levenshtein II distance can be interpreted as the count of the elements in each sequence that are not involved in the LCS and should be more sensitive to spell durations and sequencing. OM lies between these two distance measures and is the sum of two terms: a weighted sum of time shifts (indels) and a weighted sum of the mismatches (substitutions) remaining after the time shifts. High indel costs render the dissimilarity extremely time sensitive, whereas low indel costs—with respect to substitution costs—downplay the importance of time shifts in sequence comparisons. Costs also allow for state-dependent dissimilarities between sequences.

The OM distance can be thought of as based on the longest partially matched subsequence (Lesnard, 2010). From a sociological point of view, this partially matched subsequence can be interpreted as a 'common backbone' or 'common narrative' between trajectories (Elzinga and Studer, 2015).

OM has been criticized because of the lack of sociological meaning of the transformation operations, and their associated costs (Abbott and Tsay, 2000; Abbott, 2000; Levine, 2000; Wu, 2000; Aisenbrey and Fasang, 2010; Lesnard, 2010). Furthermore, the high number of indel and substitution costs may be seen as an overparameterization (Wu, 2000). Next, we examine the various methods for setting the costs.

3.3.2. Substitution costs

There are essentially three strategies when choosing substitution costs (e.g. Abbott and Tsay, (2000) and Hollister (2009)).

3.3.2.1. Theory-based costs. The first strategy is to determine the costs on theoretical grounds. *A priori* knowledge often provides an order of magnitude of the similarity of two states, which allows us to rank possible replacements. To illustrate, assume careers coded by using the following statuses: Senior Manager, S, Manager, M, and Employee, E. From the nature of the states, S is closer to M than to E. To reflect this hierarchy, we could, for instance, set the cost of replacing S with E as 1.5 times the substitution cost between S and M. In doing so, we account for the order between the states, although the exact values that are chosen for the ratios between the substitution costs remain quite arbitrary.

3.3.2.2. Costs based on state attributes. A solution that was advocated by Hollister (2009) to make the choice less arbitrary is to specify the list of state attributes on which we want to evaluate the closeness between states. For example, for professional positions we could consider the qualification required, level of responsibility and degree of precariousness, and for cohabitational statuses we could consider the events that should have been lived to reach each situation. By specifying the values of the attributes for each state, we can then derive the pairwise substitution costs from the distances between all pairs of attribute vectors. This distance could be the Euclidean distance when all attributes are numerical. More generally, in the case of nominal, ordinal and symmetric or asymmetric binary characteristics, or even in the presence of a mix of variable types, we suggest the use of the Gower (dis)similarity coefficient (Gower, 1971). Besides explicitly rendering the state comparison criteria, the approach also has the advantage of generating costs that satisfy the triangle inequality.

3.3.2.3. Data-driven costs. A third strategy is to rely on data-driven methods. Here, a popular solution is to derive the substitution costs from the observed transition rates. The idea is to assign higher costs to substituting between states when the transitions between them are rare, and a low cost when frequent transitions are observed (Rohwer and Poetter, 2005). However, deriving the substitution costs from the transition rates is questionable, as there is no reason for transition rates to reflect state similarities. For example, 'single' and 'divorced' may be seen as close states, but, by definition, we cannot switch from divorced to single. In addition, switching from single to divorced would suppose that marriage and divorce occur during the same unit of time, which is highly unlikely. Moreover, in practice, observed transition rates are generally low and the resulting substitution costs are all close to 2. Therefore, the OM distances that are based on transition rate costs produce results which are close to those obtained by using fixed state-independent costs. A solution that generates somewhat higher and more diversified transition rates is to consider the transition between the state at t and the state $q\,(>1)$ periods ahead, rather than using the transition between two consecutive time units. Whatever the time lag q, the transition-rate-based substitution does not ensure that the triangle inequality holds.

In the spirit of the work of Rousset *et al.* (2012) that was described earlier, a conceptually better approach considers the two states a and b to be close when there is a high chance that both states will be followed by a common state c, q units of time later. In other words, states a and b are close, when they share a common future. For instance, although switching between high education and high vocational school is generally unlikely, both states may be seen as similar because they both have a high probability of leading to a managerial position, and a relatively low probability of leading to joblessness. We propose to operationalize this idea by defining the substitution cost between a and b as the χ^2-distance between the cross-sectional state distributions expected q time units after the occurrence of state a and state b,

$$\gamma(a,b) = \left[\sum_{e \in \Sigma} \frac{\left\{ p(e_{+q}|a) - p(e_{+q}|b) \right\}^2}{\sum_{f \in \Sigma} p(e_{+q}|f)} \right]^{1/2}, \quad (1)$$

where $p(e_{+q}|f)$ is the probability of moving from f to e over q units of time. Using a negative q-value, we can similarly determine costs in terms of a common past.

Another method of deriving costs from data was proposed by Gauthier *et al.* (2009). This approach is an 'optimization' procedure based on methods that are used in biology (e.g. Henikoff and Henikoff (1992)). The principle is to consider two states as close—and to assign them a low substitution cost—when they tend to occur jointly in pairs of similar sequences. Similarly, the method considers them as dissimilar—and assigns them a high cost—when they rarely co-occur in pairs of similar sequences. The method works iteratively. At each step, it successively computes each cost by keeping the others unchanged and iterates until the costs converge. Experimenting with the implementation of the method in T-COFFEE (Notredame *et al.*, 2006), we faced serious issues, such as obtaining negative costs and, as a result, negative dissimilarities. Therefore, we did not include this method of computing substitution costs in our simulation study.

3.3.3. indel costs

Despite the importance of indel costs in controlling time warp, choosing indel costs has, with the noticeable exception of Stovel and Bolan (2004) and Hollister (2009), received far less attention than substitution costs. Note that Stovel and Bolan (2004) suggested lowering the indel for incomplete sequences. Such costs that change from one sequence to another would probably result in dissimilarities that violate the triangle inequality.

3.3.3.1. Single-indel cost. indel is often seen as a gap insertion operator. Thus, most applications use the same indel cost, irrespective of the inserted or deleted state. The only choice then concerns the level of this fixed indel cost. Abbott and Tsay (2000) advocated the use of a low indel cost and suggested a value in the vicinity of 0.1 times the maximum substitution cost. However, as pointed out by Hollister (2009), using such a low value 'throws out much of the careful consideration a researcher puts into creating substitution costs in the first place', because an insert and a delete would be used in place of any substitution costing more than twice the indel cost. For the extended $\mathbf{\Gamma}$-matrix to fulfil the triangle inequality and if we want indels to serve only to adjust sequence lengths, a unique indel cost c_{I} should be within the range $\gamma_{\max}/2 \leqslant c_{\mathrm{I}} \leqslant L\gamma_{\max}/2$, where $\gamma_{\max}$ is the maximum substitution cost and L is the maximum sequence length.

3.3.3.2. State-dependent indel costs. Little attention has been paid to state-dependent indel costs. According to Stovel (2001), more exceptional or rare states should be given a higher cost. Like the resemblance between states, we can determine how exceptional a state is, theoretically, on the basis of the state's attributes, or from the data. As a data-driven solution, we propose to define the indel cost of state a as a monotonic function—such as a logarithm or square root—of the inverse of the overall observed frequency of state a or, equivalently, of the inverse mean time spent in state a. An alternative could be to use the mean time not spent in a. Such data-driven solutions for indel costs avoid the criticisms of transition-rate-based substitution costs. An alternative to the latter method could be to set substitution costs as the sum of the indels of the two terms involved.

3.4. Variants of optimal matching

Despite the high flexibility of OM with state-dependent costs, several researchers (Elzinga, 2003; Hollister, 2009; Halpin, 2010; Elzinga and Studer, 2015) have pointed out that OM distances are essentially driven by differences in durations. There are two main reasons for this. First, sequences in social sciences typically comprise a few long spells. Therefore, the LCSs typically include these longest spells or long portions of them (Elzinga and Studer, 2015). Second, OM operations are independently applied on each symbol in the sequence, regardless of the context. OM weights the insertion of state a in sequence aa and in sequence bb equally. However, in the first case, the insertion affects only the time that is spent in the spell in state a, whereas, in the second case, it changes the sequencing (Hollister, 2009; Halpin, 2010). Lesnard (2010) observed that OM does not consider the position (i.e. the age or date) when transformation operations are applied.

The OM variants that are discussed below aim to make edit operations more context sensitive by making them depend either on the position in the sequence where the operation applies or on the surrounding patterns at that position.

3.4.1. Dynamic Hamming distance

State similarities in time-use analyses—e.g. between sleeping and commuting—can hardly be assumed to remain the same all day, and distinct timings reflect important differences in behaviour. As a result, Lesnard (2010) focused on OM without indels, such as the generalized Hamming distance, and proposed that substitution costs should depend on the position t in the sequence. He operationalized the idea by deriving the substitution cost at t from the cross-section of the transition rates observed between $t-1$ and t and between t and $t+1$.

The dynamic Hamming distance DHD shares the strong timing sensitivity of the Hamming distance. Several criticisms can be pointed out. First, the criticism of the validity of the transition-rate-based substitution costs applies here also. Second, the number of transition rates to estimate is very high, potentially worsening overparameterization. Furthermore, if the meaning of a state a changes with the time when it occurs, a simpler solution could be to consider state a at time t and state a at time $t' \neq t$ as two distinct states a_t and $a_{t'}$.

3.4.2. Localized optimal matching

The OM extension that was proposed by Hollister (2009) aims to make indel costs dependent on the two adjacent states. The motivation is that inserting or deleting a state that is similar to its neighbours would change only the length of the spell in that state, without affecting the sequencing. However, an indel of a state that is different from its neighbours has much more important consequences and should, therefore, be charged a higher cost.

This localized OM is controlled with two user-defined parameters. The first, e, can be interpreted as a spell expansion cost or time warp penalization. The second parameter, g, penalizes differences with surrounding states measured by the substitution costs. (For indels at one of the sequence ends, the average between the costs of the substitutions with the two surrounding states is replaced by the cost of the substitution with the sole adjacent term.) In her experiments, Hollister (2009) obtained the best results with a small shift penalization e and a g close to $1-2e$. As long as parameters e and g fulfil the constraint $1-2e \leqslant g$, the method also prevents the OM from using a pair of indels instead of a substitution. Thus, it provides a way to allow for important time warps while preserving the effectiveness of substitution costs.

By construction, the localized OM should be less sensitive than the classical OM to differences

in spell length, while being more sensitive to changes in sequencing. However, the localized OM can generate dissimilarities that do not satisfy the triangle inequality (Halpin, 2014).

3.4.3. *Optimal matching sensitive to spell length*

The OM sensitive to spell length variant, proposed by Halpin (2010), accounts more explicitly for the spell length, making indel and substitution costs depend on the spell length. Operations inside longer spells cost less than those involving shorter spells. More precisely, the costs are multiplied by a factor of $1/t^h$, where t is the spell length and $0 \leqslant h \leqslant 1$ the exponent time weight.

Decreasing the indel cost with the spell length produces the expected effect of favouring indels in longer spells instead of, for instance, indels that would create or suppress spells. However, the decrease in the substitution costs with the lengths of the implied spells has the reverse effect of encouraging the splitting of long spells. These contradicting effects make it difficult to predict the sensitivity of the measure to spell lengths. Moreover, this dissimilarity does not guarantee that the triangle inequality holds (Halpin, 2014).

3.4.4. *Optimal matching between sequences of spells*

To overcome the limitations of the two previous context-sensitive dissimilarities, we propose to measure the OM distance between sequences of spells. The general idea is to consider, for each value of t, a spell in state a during t units of time as a distinct element, denoted a_t, of the alphabet. Doing so considerably increases the size of the alphabet and, as a consequence, the number of indel and substitution costs to be considered. However, the number of parameters can easily be limited if we express the cost $c_{\mathrm{I}}^{\mathrm{S}}(a_t)$ of the indel of spell a_t and the substitution cost $\gamma^{\mathrm{S}}(a_{t_1}, b_{t_2})$ between spells, a_{t_1} and b_{t_2} respectively, in terms of the basic indel and substitution costs ($c_{\mathrm{I}}(a)$ and $\gamma(a,b)$) of the constituent elements, a and b, and a correction factor function of the spell length. For instance, letting $\delta \geqslant 0$ be a weight factor for the spell length, the costs can be defined as

$$c_{\mathrm{I}}^{\mathrm{S}}(a_t) = c_{\mathrm{I}}(a) + \delta(t-1), \tag{2}$$

$$\gamma^{\mathrm{S}}(a_{t_1}, b_{t_2}) = \begin{cases} \delta\,|t_1 - t_2| & \text{if } a = b, \\ \gamma(a,b) + \delta(t_1 + t_2 - 2) & \text{otherwise.} \end{cases} \tag{3}$$

The parameter δ is the cost of extending or compressing a sequence by 1 unit of time, and the substitution between two spells $\gamma^{\mathrm{S}}(a_{t_1}, b_{t_2})$ is the cost of compressing each spell into a 1-unit-long spell, plus the substitution between the two states concerned, a and b. For $\delta < c_{\mathrm{I}}(a)$, inserting an a in an existing spell a_t costs less than creating a new spell in a. Therefore, the method favours the expansion (or compression) of existing spells. However, unlike the method of Halpin, it does not encourage breaking long spells. Moreover, as defined by equations (2) and (3), the costs $c_{\mathrm{I}}^{\mathrm{S}}(\cdot)$ and $\gamma^{\mathrm{S}}(\cdot)$ satisfy the triangle inequality as long as $c_{\mathrm{I}}(\cdot)$ and $\gamma(\cdot)$ satisfy the inequality. Interestingly, for $\delta = 0$, the OM of spell sequences becomes the OM distance between the DSS sequences.

The OM between sequences of spells is, by construction, sensitive to differences in the duration of spells. It is also sensitive to sequencing by considering the DSS sequence and allows some control for the time warp through the expansion–compression penalty factor δ.

3.4.5. *Optimal matching between sequences of transitions*

Another way of accounting for the context, as described by Biemann (2011), is to compute

the OM distances between sequences of transitions. The transitions in a state sequence are characterized by the two successive long subsequences obtained by joining each state with its previous state. For example, the transitions in *aabb* are *aa–ab–bb*. We could possibly also specify the start of a sequence by using a transition from the start to the first state and, likewise, the end of the sequence by using a transition to the end state.

As noted by Biemann (2011), by considering transitions instead of the states, we increase the size of the alphabet considerably and, hence, the number of indel and substitution costs to be considered. To overcome this limitation, we propose (similarly to the case of sequences of spells) to express the indel $c_{\mathrm{I}}^{\mathrm{B}}(a \to b)$ and substitution $\gamma^{\mathrm{B}}(a \to b, c \to d)$ costs of transitions in terms of the indel and substitution costs of states. Considering that a transition $a \to b$ comprises an origin state a and a type of transition (e.g. a transition to the same state or a transition to another state), we express the cost of inserting or substituting a transition as a linear combination of respectively the cost of inserting or substituting the origin state and the cost $c_{\mathrm{T}}(a \to b)$ of the transition type concerned. Formally, we define the indel and substitution costs as follows:

$$c_{\mathrm{I}}^{\mathrm{B}}(a \to b) = w c_{\mathrm{I}}(a) + (1-w) c_{\mathrm{T}}(a \to b), \tag{4}$$

$$\gamma^{\mathrm{B}}(a \to b, c \to d) = w\gamma(a,c) + (1-w)\left\{c_{\mathrm{T}}(a \to b) + c_{\mathrm{T}}(c \to d)\right\}, \tag{5}$$

with $c_{\mathrm{I}}(a)$ the (possibly normalized) indel cost of the origin state a, $c_{\mathrm{T}}(a \to b)$ the transition type cost, $\gamma(a,c)$ the (possibly normalized) substitution cost between the origin states, a and c, and $w \in [0,1]$ a coefficient that controls the trade-off between the cost that is related to the origin state and the cost that is related to the type of transition. A simple parameter-free solution for the $c_{\mathrm{T}}(a \to b)$ function is to set it to 0 when $a = b$, and 1 otherwise. An alternative, which would make $c_{\mathrm{T}}(a \to b)$ state dependent without the need for any additional parameters, is to set $c_{\mathrm{T}}(a \to b)$ as the substitution cost $\gamma(a,b)$ between a and b. Both solutions generate costs $c_{\mathrm{I}}^{\mathrm{B}}(\cdot)$ and $\gamma^{\mathrm{B}}(\cdot)$ that satisfy the triangle inequality when the basic costs $c_{\mathrm{I}}(\cdot)$ and $\gamma(\cdot)$ themselves verify the inequality.

The OM of sequences of transitions is, by construction, sensitive to differences in sequencing. With our formulation of the indel and substitution costs of the transitions, we obtain the classical OM for $w = 1$, which shares the properties of OM. Otherwise, by reducing w, we can increase the sensitivity to sequencing. Time warp can be controlled through the origin state indel cost $c_{\mathrm{I}}(a)$.

Finally, it is worth mentioning that Dijkstra and Taris (1995) proposed an interesting distance measure that should account for some special aspects such as the number and the order of common states. However, as shown by van Driel and Oosterveld (2001), their algorithm does not produce the expected results.

4. Simulation study

So far, we have reviewed a great number of dissimilarity measures between sequences. Moreover, many dissimilarity measures depend on user-defined parameter values and, thus, define families of measures. However, in the end, we face the crucial question of choosing between them.

To help in that choice, this section provides empirical insights on how dissimilarity measures behave with regard to the three aspects that are relevant to comparing state sequences that describe life trajectories, namely sequencing, duration and timing. In what follows, we report the main results from a series of simulation strands.

The simulations reported provide an original view of the ability of the dissimilarity measures to render differences in each of the sequencing, duration and timing dimensions. In this regard,

our simulations differ from other attempts to compare dissimilarity measures empirically. Several researchers (for a review, see Halpin (2014) and Robette and Bry (2012)) have analysed how results—most often the clusters that are derived from the dissimilarity values—change with each dissimilarity measure used. Such approaches permit us to assess the robustness of the outcome of the dissimilarity-based analyses against the dissimilarity measure that is used. However, outcome-oriented simulation analyses do not *in stricto sensu* provide indications on the behaviour of the measures, and the generalization of their findings to other data sets and analysis methods—clustering algorithms—is subject to debate. The approach by Robette and Bry (2012), which is based on correlations between dissimilarities computed on artificial data, is more illuminating from that point of view. Nevertheless, although the Mantel tests of the correlations that were used by Robette and Bry prove useful in identifying measures that behave similarly, they do not specify what the measures are sensitive to.

4.1. Simulation design

The simulation study consists of different strands, each of which studies the sensitivity of the dissimilarity measures to one of timing, duration or sequencing. Each strand may itself contain a series of simulations run with different specifications of the differences tested.

The general principle of each series of simulations is to generate, in a controlled manner, two groups of sequences that differ in a selected *single aspect* of interest. In each group, the characteristic evaluated—sequencing, duration or timing—is kept fixed for all sequences in the group, whereas the other aspects are changed randomly across the sequences to allow for non-systematic differences between sequences on these other non-evaluated aspects. Doing so, the sequences compared differ systematically in the aspect evaluated, but also differ randomly on all other aspects. Thus, we can evaluate the relative importance that is given by the dissimilarity measures to the selected aspect in the presence of discrepancies on the others. Hence, we can evaluate how well each measure renders differences of the aspect studied.

Let us illustrate with an example. To measure the sensitivity to sequencing, we generate two groups of sequences with a different unique sequencing pattern in each group. Whereas the order of the states remains identical for all sequences inside a group, the timing and time spent in each distinct successive state are changed randomly within the groups. Therefore, a dissimilarity measure that is more sensitive to differences in duration than in sequencing will probably take similar values for pairs of sequences belonging to the same group as it would for pairs with a sequence from each of the two groups. In contrast, a measure that is sensitive to sequencing will typically take higher values for dissimilarities between groups than it would for dissimilarities within groups.

The sensitivity to the criterion considered is measured with the pseudo-R^2 that was defined in Studer *et al.* (2011), which measures the proportion of the discrepancy of the sequences explained by a categorical covariate. In our case, the covariate is the two-group variable. The discrepancy of the sequences is evaluated from the pairwise dissimilarities in the same way as the variance of a series of values can be derived from the pairwise differences between the observed values. A high R^2-value will reflect strong sensitivity to the considered systematic difference between the two groups. In other words, a high R^2 means that the measure can discriminate between the groups. In contrast, a low R^2-value indicates that the measure poorly accounts for the tested dimension. To ensure stable R^2-values, we generate 1 million sequences per group in each series of simulations. All simulated sequences are of length 20. Despite the huge number of sequences, all computations could be done relatively quickly by considering only unique sequences and weighting them by their counts (Studer, 2013). Each set of 1 million simulated

Table 2. Designs for evaluating sensitivity to order, timing and duration of states

Tested dimension	*Description*	*Group 1*	*Group 2*
Sequencing	Order patterns controlled in each group, and duration in each consecutive state left random under the constraint of the fixed sequence length	'*abc*' '*abca*' '*abcda*' '*abca*' '*abab*' '*abc*' '*abc*' '*abcd*'	'*cba*' '*acba*' '*adcba*' '*abda*' '*baba*' '*abd*' '*acb*' '*cdab*'
Timing	Sequences randomly follow one of the patterns '*abcde*' or '*edcba*', and the time point t that the spell in state '*c*' must cover is controlled	$t=7$ $t=15$	$t \in \{9 \ldots 15\}$ $t \in \{7 \ldots 13\}$
Duration	Sequences randomly follow one of the patterns '*abc*' or '*cba*' and duration d of the spell in state '*b*' is controlled	$d=4$ $d=14$	$d \in \{6 \ldots 14\}$ $d \in \{4 \ldots 12\}$

sequences contained between 80 and 800 unique sequences, except for one set that had around 3000 unique sequences.

For each series of simulations, we obtain an R^2_d for each considered dissimilarity measure d. The R^2_d-values can be compared across dissimilarity measures within each series, where all R^2_d are computed on the same set of sequences. However, they are not comparable between series or strands, since the total variability and the mean R^2 differ significantly across series. Therefore, we report the standardized value, namely the score. This score reflects the sensitivity of each dissimilarity measure d in comparison with the overall sensitivity of all measures considered. The score is positive for dissimilarity measures that are more sensitive than the average to the tested dimension, and negative otherwise.

4.2. Random-sequence generation

We ran two sets of simulations, each with a different sequence-generating model. For the first set, we generated the *state sequences* directly. For the second set, we postulated assumptions on the *occurrences of events* and then derived the states from these events.

For clarity, and for brevity, we report only a subset of the series of simulations that we tried (see Studer (2012)). However, the experiments that are reported are representative in that they render all salient findings of the complete set of simulations.

4.2.1. State-based generating process

The direct generating process is based on the duration-stamped spell representation of sequences. It first determines the sequence $\mathbf{x} = (x_1, \ldots, x_{l_{\text{dss}}})$ of the l_{dss} DSS, and then the durations $\mathbf{t} = (t_1, \ldots, t_{l_{\text{dss}}})$ of the successive distinct states. The order—the DSS—is chosen randomly from a list of possible sequencings, and the durations are set randomly assuming uniform distributions. For each series, the control of the aspect tested is achieved by means of constraints on the generating process.

We report three strands of state-based simulations, the characteristics of which are summarized in Table 2. Each strand comprises several series of simulations. The first strand evaluates the sensitivity to sequencing by completely controlling the order in each of the two groups in each series.

Table 3. Designs for evaluating sensitivity to a random change of state

Description	*Order pattern*	*State change in group 2*
Controlled order pattern and random durations:	'*abc*'	Anywhere
sequences in the second group derived from the	'*abc*' or '*cba*'	Anywhere
sequences of the first group by randomly	'*abc*'	Start or end of spells
changing one of their elements	'*abc*' or '*cba*'	Start or end of spells

Table 4. Simulations evaluating the sensitivity to the order and timing of events, and to duration between events

Simulation	*Description*	*Group 1*	*Group 2*
Order	Random-occurrence times	$e_1 < e_2$	$e_1 > e_2$
Timing	Date e_1 of event 1 is fixed	$e_1 = 4$	$e_2 \in \{6 \ldots 14\}$
		$e_1 = 14$	$e_2 \in \{4 \ldots 12\}$
Duration	Fixed duration between events	$e_2 - e_1 = 2$	$e_2 - e_1 \in \{4 \ldots 12\}$
	e_1 and e_2	$e_2 - e_1 = 12$	$e_2 - e_1 \in \{2 \ldots 10\}$

The second strand evaluates the sensitivity to timing. The order patterns are selected randomly and the durations are set randomly, while controlling the start of the spell in state c in each of the two groups. Several series of simulations are run with varying differences in the time point where we impose to be in state c between the two groups. Finally, the third strand evaluates the sensitivity to duration by controlling the total consecutive amount of time spent in a given state for each group.

We ran an additional strand of simulations to evaluate the sensitivity of the measures to small perturbations (Table 3). The same order is retained for the two groups but, in group 2, the sequences are perturbed by randomly changing the state of an element in the sequence, either for any element or for one element among those at the junction of two successive spells.

4.2.2. Event-based generating process

The aim of this second group of simulations is to evaluate the sensitivity of the measures to the underlying events that provoke the change in states.

We consider the occurrences of successive events and define the consequent new state after each event. In our simulations, we consider three events. Each sequence is then characterized by when each of the events e_1, e_2 and e_3 occurs. The sequences are simulated by generating the times of occurrence with an independent uniform distribution over the period of observation for each of the three events.

Three strands of event-based simulations are considered. The first evaluates the sensitivity to the order that events occur. Here, we impose the restriction that the first event occurs before the second event in group 1, and after the second event in group 2. The second strand evaluates the sensitivity to the timing of the events by controlling the occurrence of event 1 in each group. To evaluate the sensitivity to the duration between two events, in the last strand, we control the elapsed time between the first two events. The time of occurrence of the third event is left as random in all cases. Table 4 summarizes the simulations.

Table 5. Distances included in the simulation study

Distance	*Configurations*
Distribution based	EUCLID($K=1$) (Euclidean), CHI2($K=1,2,4,5,10,20$), (χ^2-distance between distributions within K periods), CHI2fut (metric based on distributions of subsequent states)
Hamming	HAM (simple and generalized Hamming) DHD (dynamic Hamming)
OM	OM, OM($i=1.5$), OM(trate), OM(indelslog), OM(indels), OM(future)
Localized OM (OMloc)	OMloc($e=0,0.1,0.25,0.4$)
Spell-length-sensitive OM (OMslen)	OMslen($h=1, i=1,1.5,5$), OMslen($i=1,1.5,5$)
OM of spell sequences (OMspell)	OMspell($e=0,0.1,0.5,1$), OMspell($e=0,0.1,0.5,1, i=2$)
OM of transition sequences (OMstran)	OMstran($w=0,0.1,0.5$), OMstran($i=1.5, w=0.1,0.5$), OMstran($i=5$, tm=sm, $w=0.1,0.5$), OMstran(tm=raw)
Number of matching subsequences (NMS)	NMS
Subsequence vectorial representation (SVRspell)	SVRspell($b=0,1,2,3$), SVRspell($b=0,1,2,3, a=1$)

Table 6. Meaning of parameters

Label	*Description*
K	Number of intervals used to compute χ^2- and Euclidean distances
i	indel cost (single cost of 1 when not specified)
sm†	Substitution cost (single cost of 2 when not specified): trate (derived from transition rates), indelslog (derived from log-state-frequency-based indel costs), indels (derived from inverse state-frequency-based indel costs), future (common future), *ec* (based on count of non-shared experienced events)
e	Spell expansion cost (for OMspell and OMloc)
w	Weight of origin state *versus* transition-type trade-off
ti	Transition indel costs (single cost of 1 when not specified): sm (based on substitution costs), raw (Biemann's method)
a	Subsequence length weight exponent (0 when not specified)
b, h	Spell duration weight exponent for SVRspell and OMslen respectively (when not specified, $b=1$ and $h=0.5$)

†For brevity 'sm=' will be omitted and therefore OM arguments without the '=' sign should be interpreted as values of the sm argument.

When comparing event-based sequences, we can, for state-dependent measures, define the state dissimilarities—substitution costs—by using the number of unshared underlying events. For example, the substitution cost between state 'has experienced event e_2 only' and state 'has experienced all three events' is 2, since two events distinguish these states. We use this principle to test the behaviour of measures parameterized with features-based costs.

4.3. Analysed dissimilarity measures

Most of the dissimilarity measures that are described in Table 1 have been included in the simulation study. For distances that can be parameterized, we consider a selection of parameter configurations to explore the effect of the parameters and the range of behaviour that can be covered. The complete list of dissimilarity measures and parameter configurations studied in the simulations is given in Table 5. The meanings of the parameters that are shown in Table 5 and in the following figures are specified in Table 6.

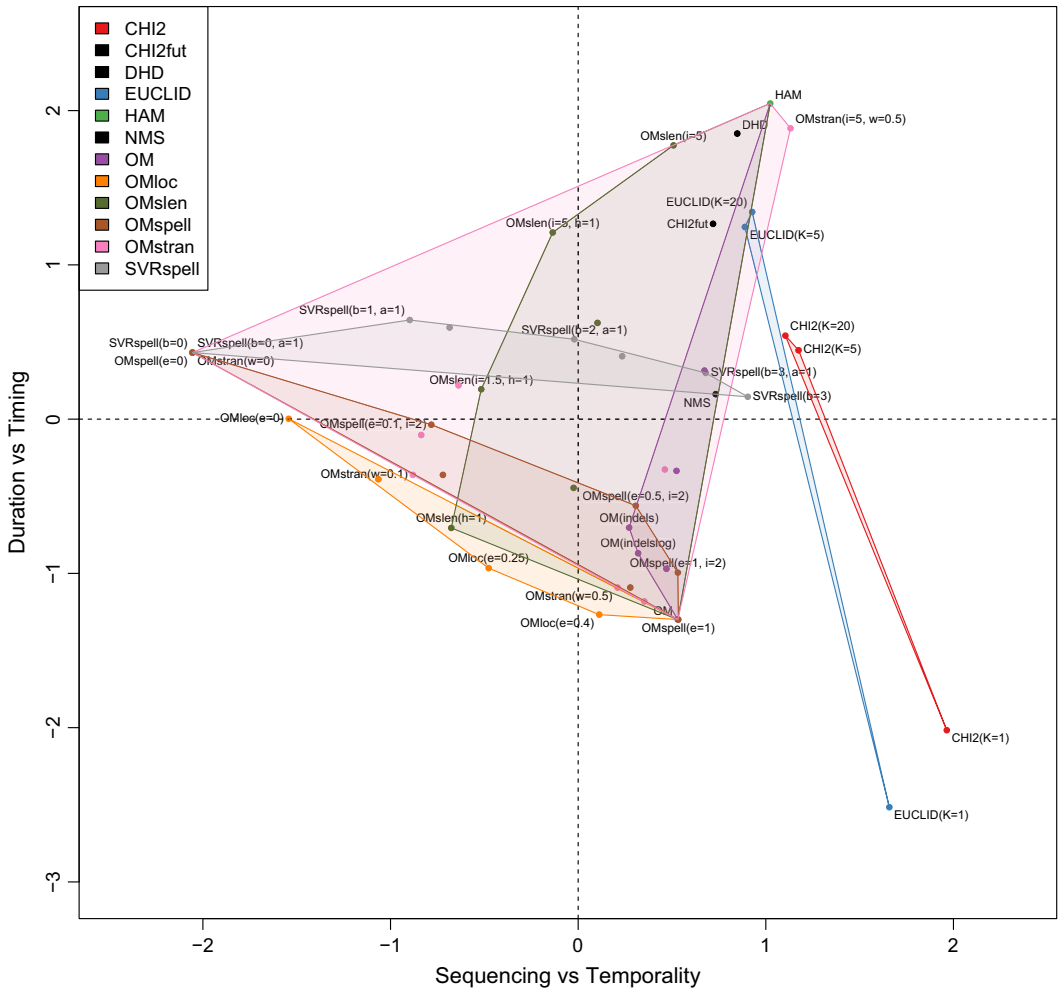

Fig. 1. Scores for state-based simulations

4.4. Results

Detailed results for each series of simulations are provided as an on-line appendix. Here, we summarize the outcome of the study by opposing the mean scores that are achieved by each measure for the simulations of the 'sequencing' strand to the mean scores that are obtained for the temporality—'timing' and 'duration'—strands on one side, and the duration scores to the timing scores on the other. (The mean temporality scores are computed as the average between the mean timing and mean duration scores, and the mean scores that are reported have been standardized.) The duration and timing axes roughly correspond to the first two robust principal components (Todorov and Filzmoser, 2009) that were found in Studer (2012). However, unlike principal components, the axes here are defined independently from the data. They also have a clearer interpretation. The first axis is defined as the temporality score minus the mean sequencing score. Therefore, it is oriented such that higher sensitivity to sequencing appears on the left and higher sensitivity to temporality dimensions appears on the right. The second axis is defined with higher sensitivity to durations at the bottom and higher sensitivity to timing at the top.

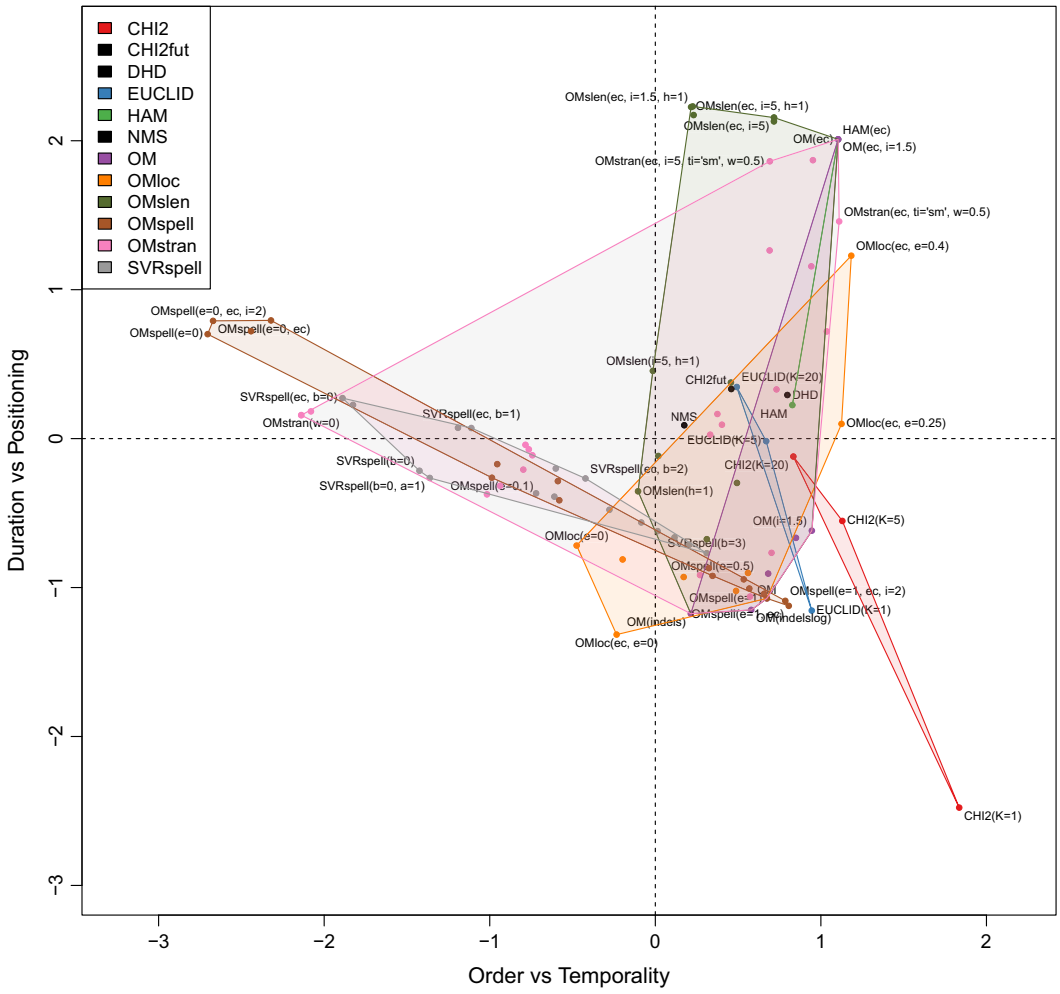

Fig. 2. Scores for event-based simulations

Results from the state-based group of simulations are displayed graphically in Fig. 1 and the results for the event-based group are shown in Fig. 2. Fig. 3 reports the results for sensitivity to a random change of one token in the sequence. In each figure, the position of the measures should be interpreted relatively to the others and does not reflect an absolute level of sensitivity.

In order not to overload Figs 1–3 with too many points, the results for each family of measures is represented by the smallest polytope covering the scores that were obtained for the various parameterizations tested. The labels of inner points have been omitted and only those of configurations that are associated with the vertices of the polytope are displayed. A large polytope area, such as that for OMstran—OM of transitions—in Fig. 1 indicates that the measure allows for very different sensitivities through its parameterization.

We can observe that the measures are distributed within a triangle in Figs 1 and 2. This (unsurprisingly) reflects a higher contrast between duration and timing sensitivities between measures that are sensitive to temporal aspects—on the right—than among measures that are primarily sensitive to the sequencing—on the left. A noticeable general outcome in Fig. 2 is that

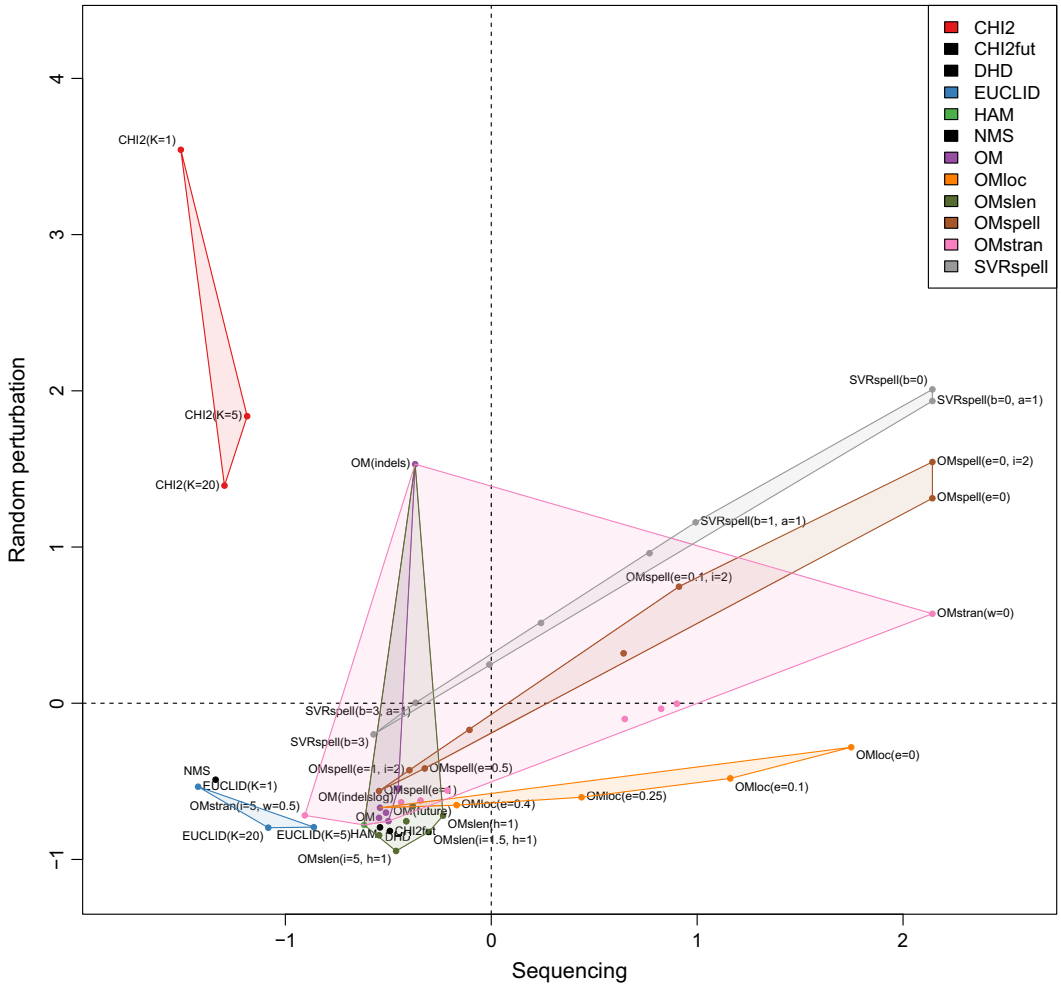

Fig. 3. Sensitivity to a random change of state *versus* sensitivity to sequencing

considering explicit information on the state proximities can significantly affect the behaviour of the measure (e.g. HAM(*ec*) lies far from HAM).

4.4.1. Results by distance families

Here, we examine each considered family of dissimilarity measures in more detail. Figs 4(a)–12(a) and Figs 4(b)–12(b) respectively give the position that each family occupies in Figs 1 and 2.

4.4.1.1 Distribution-based distances. Unsurprisingly, dissimilarity measures based on differences between distributions over the whole period ($K = 1$) appear to be the most sensitive to durations (Fig. 4). They are also the least sensitive to differences in sequencing, with R^2s close to 0. The χ^2-distance CHI2($K = 1$) is, among all distances considered, the most sensitive to duration differences for rare states, whereas the Euclidean distance EUCLID($K = 1$) is the most

sensitive to differences for states with high durations. When K increases, the sensitivity of the χ^2- measure shifts from duration to timing. For K equal to the sequence length (here, 20), CHI2 receives scores that are similar to those of the Hamming family regarding timing but maintains some sensitivity to differences in durations. The detailed results in the on-line appendix show that the positionwise χ^2-distance ranks better as a time-sensitive measure for small time changes than for large differences in timing. Here, CHI2fut—itself a positionwise measure—is closer to the positionwise CHI2 than are CHI2-versions with a smaller K.

4.4.1.2. Hamming. All variants of the Hamming distance lie in the top right-hand quadrant, meaning that they are specifically sensitive to timing differences (Fig. 5). They are slightly less insensitive to differences in sequencing than overall distribution-based distances. This is because sequencing is partly determined from the start and end states, especially when, as in our generated sequences, the number of transitions remains low. The neutral position of HAM and DHD on the vertical timing–duration scale for the event-based sequences is a consequence of the much higher timing sensitivity that is reached by HAM(*ec*) by using event-based substitution costs. The HAM- and DHD-scores are low relative to the HAM(*ec*)-score. From the simulations, the time varying costs of the DHD-metric seem to relax the strong time sensitivity of pure Hamming distance somewhat. As with the positionwise χ^2-distances, the Hamming distances rank better as time-sensitive measures for small time changes than for large time differences.

4.4.1.3. Optimal matching. The family of OM distances lies on the right of the plot, which confirms the low sensitivity to differences in sequencing, as pointed out, for instance, by Elzinga (2003), Hollister (2009) and Halpin (2010) (Fig. 6). As expected, we also observe that OM with high indel costs—relative to substitution costs—is more sensitive to timing differences (remember that HAM is OM with an arbitrarily high indel cost). Further, lowering the indel cost increases the sensitivity to duration and seems to reduce the insensitivity to sequencing at the same time. As expected, the scores for OM with data-driven substitution costs remain very close to those with a single state-independent cost of 2, the variation in position being essentially determined by the ratio between indel and substitution costs. For example, the data-driven costs of OM(future)—the unlabelled point inside the OM area slightly below the vertical splitting line—are low in comparison with those used in other OM versions, which, for the same indel value, increase the indel/substitution cost ratio. This explains why OM(future) lies higher in the plot. As also expected, deriving substitution and indel costs from the state frequencies renders OM more sensitive to changes in the duration of rare events (see Fig. 3). Using the logarithm of inverse frequencies seems a better choice than raw inverse frequencies that make OM too sensitive to rare events and small perturbations. Costs that are derived from the count of non-shared lived events, *ec*, reduce the sensitivity to duration and ensure that more importance is placed on timing differences. Interestingly, in all our simulations, OM(*ec*) and OM($ec, i = 1.5$) receive the same scores as HAM(*ec*). The reason for this is that the substitution costs are so low globally that indels costing 1 or more are never used.

4.4.1.4. Localized optimal matching. The distances of the localized OM family are, with a few exceptions in the case of costs based on the count of non-shared events, in the lower portion of the plots (Fig. 7). The horizontal position is determined by the expansion cost e. In this case, lower values of e mean that the position is further to the left, with the measure becoming highly insensitive to temporality as e approaches 0. For some simulation strands, the part of R^2 for timing differences that is accounted for by the measure becomes negative. This means that, with

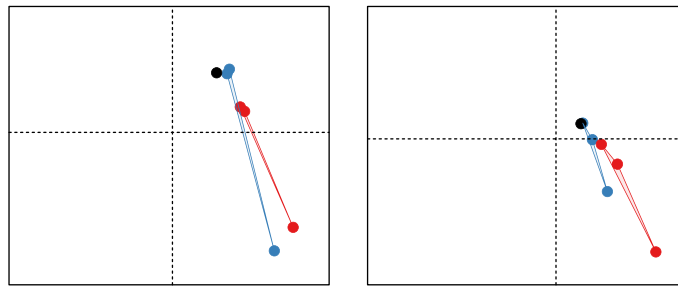

Fig. 4. Distribution-based distances: (a) state; (b) event

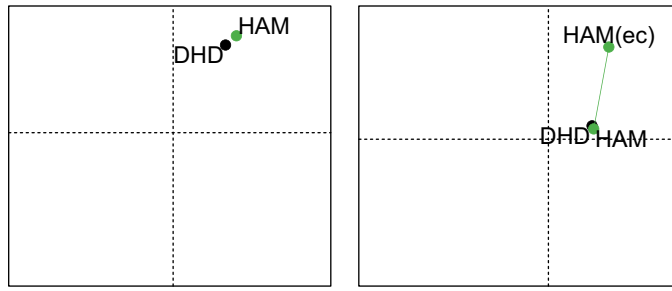

Fig. 5. Hamming distance: (a) state; (b) event

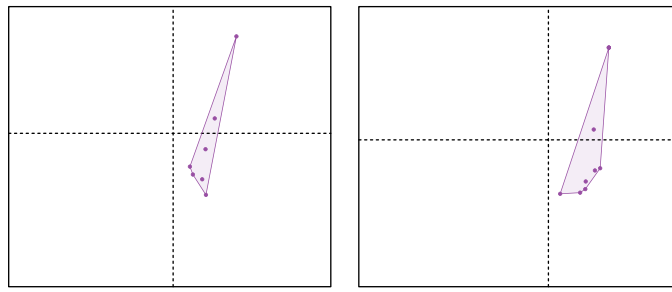

Fig. 6. OM: (a) state; (b) event

OMloc, we can obtain a total within-group discrepancy that is greater than the overall 'OMloc'-discrepancy. This is a consequence of the violation of the triangle inequality and makes OMloc especially unsuited to distinguishing between groups of sequences with different timings.

4.4.1.5. Spell-length-sensitive optimal matching. As expected by Halpin (2010), OMslen appears to be less sensitive to differences in durations than classic OM (Fig. 8). However, here again, for several simulation strands (related to timing, duration and random perturbation) we obtained negative R^2s when $h = 1$.

4.4.1.6. Optimal matching of spells. The family of OMspell distances is around a line going from right to left from OM (i.e. classic OM with a single substitution cost) to OM of the DSS

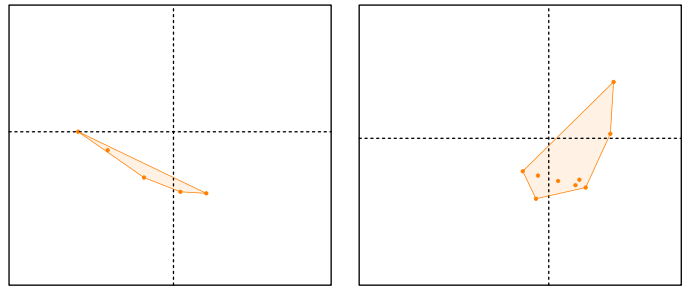

Fig. 7. Localized OM: (a) state; (b) event

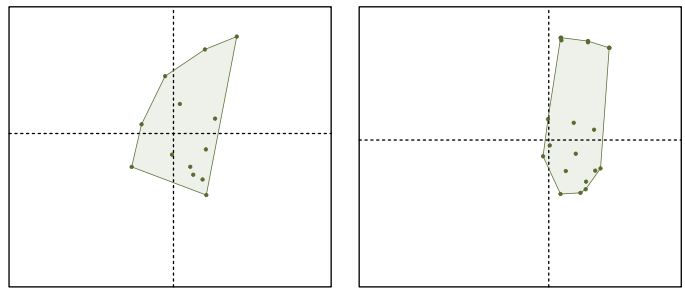

Fig. 8. Spell-length-sensitive OM: (a) state; (b) event

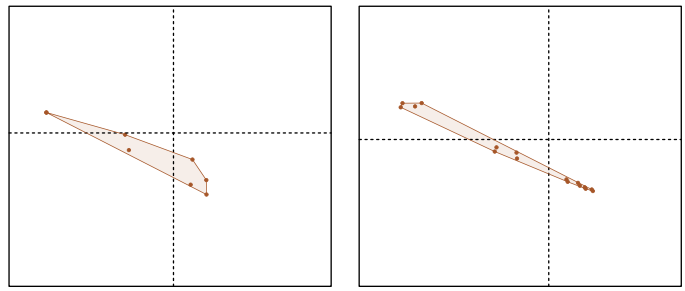

Fig. 9. OM of spells: (a) state; (b) event

sequences, the latter corresponding to OMspell ($e = 0$) (Fig. 9). A high expansion cost e makes the measure more sensitive to temporality, whereas low values of e give more importance to sequencing. The sensitivity to temporality is attributable primarily to duration rather than to timing.

4.4.1.7. Optimal matching of transitions. The family of OM distances between sequences of transitions covers the largest range of sensitivity combinations (Fig. 10). Sensitivity to temporality increases with the value of the origin–transition trade-off parameter w. Recall that, for $w = 1$, OMstran is equivalent to classic OM. Lowering the value of w significantly increases the sensitivity to sequencing. As with classic OM, the vertical position is driven mainly by the indel/substitution cost ratio. However, we can observe that the effect of the ratio becomes smaller

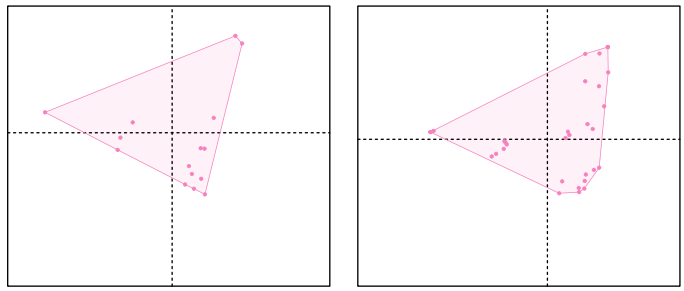

Fig. 10. OM of transitions: (a) state; (b) event

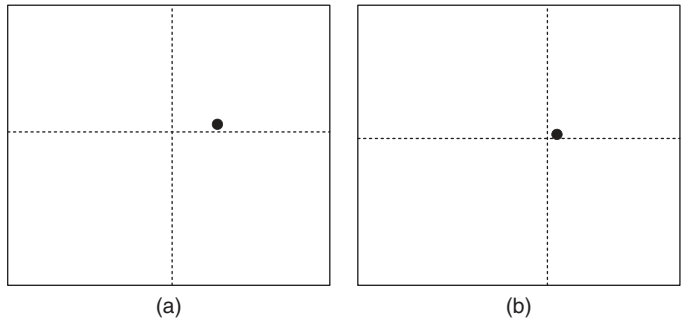

Fig. 11. NMS: (a) state; (b) event

as w decreases; in other words, when more importance is given to the transition type than to the origin states.

4.4.1.8. NMS. The NMS-distance occupies a neutral position near the centre of the plots (Fig. 11). Such neutral positions result in other distances exhibiting balanced positive sensitivity to sequencing, temporality and duration. However, this is not so for NMS, which appears to be virtually insensitive to each of them. For instance, in Fig. 3, we observe that NMS is the least sensitive distance to ordering. Counterintuitively, the measure receives its best scores for sensitivity to timing. This strange behaviour is a consequence of the extremely low proportion of subsequences that match among the huge number of subsequences in each sequence. The NMS-measure exhibits the expected sensitivity to sequencing when applied to the DSS sequences, which corresponds to SVRspell ($b = 0$).

4.4.1.9. SVRspell. The family of SVRspell-distances—also based on matching subsequences—does not suffer from the NMS lack of sensitivity to our three relevant aspects (Fig. 12). The position of the measure is essentially linked to the spell duration exponent weight, b. For $b = 0$ and $a = 0$, the SVRspell distance becomes the NMS-distance between the DSS sequences. This is the configuration that is the most sensitive to sequencing. Increasing b makes the measure more sensitive to temporality. Overall, and contrary to what we expected, SVRspell lies in the top half of the state-based simulation plot and, therefore, looks more sensitive to timing differences than to differences in durations. However, this behaviour is not confirmed by the event-based simulations. The effect of the a-parameter that determines the weight that is given to the length of the matching subsequences remains unclear, but limited. From Fig. 3,

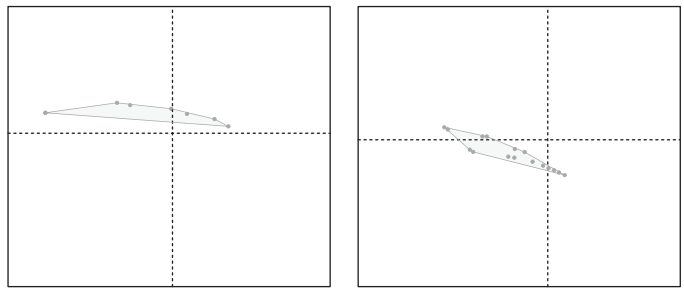

Fig. 12. SVRspell: (a) state; (b) event

we observe that the SVRspell-measures are, after the χ^2-measures, the most sensitive to small random perturbations.

4.4.2. Small random perturbations.
The sensitivity to small random perturbations is shown in Fig. 3, where the scores for a random change of one token in each sequence are plotted against sequencing scores. We observe that the basic Euclidean and χ^2-distances are, regardless of the breakdown of the covered time interval into periods, quite insensitive to differences in ordering. The χ^2-distance appears to be, in our simulations, much more sensitive to small perturbations than the basic Euclidean distance is. If we exclude the χ^2-distance, we observe that parameterizations that make measures more sensitive to ordering at the same time render them more sensitive to small perturbations. The linear correlation between the ordering scores and the scores for random perturbation of all except χ^2-measures is 0.8.

5. Choosing the right dissimilarity measure

The aim of this section is to provide guidelines on choosing from among the many different possibilities of measuring dissimilarity between sequences. The choice is difficult because it typically is multicriterial. For instance, in the retained social science framework, we expect the measure to reflect differences in timing, duration and sequencing. From our theoretical knowledge and empirical evidence on the behaviour of the various dissimilarity measures, there is no measure that dominates all others in all three dimensions of interest.

However, some distance measures are not recommended, on the basis of our study. These are the NMS-distance, OMloc (the localized OM) and OMslen (the spell-length-sensitive OM). NMS lacks sensitivity to all three aspects, OMloc has strange behaviour resulting from the violation of the triangle inequality and OMslen has counterintuitive and, hence, unexpected behaviour. Although these three measures have interesting characteristics, there are alternatives—SVR for NMS, OMstran for OMloc and OMspell for OMslen—that share similar aims without suffering from their drawbacks. Further, remember that we discarded the so-called 'optimization cost' method that was proposed by Gauthier *et al.* (2009) because of serious mathematical problems that could lead to negative dissimilarities.

Deriving substitution costs from transition rates—OM(trate)—adds complications and could possibly generate violations of the triangle inequality. Moreover, as shown by the simulations, OM(trate) provides results that are very close to those of OM with a single state-independent

substitution cost. The same holds for DHD, which produces results that are similar to the simple Hamming distance HAM. In contrast, when states are structurally organized, as in our event-based simulations, taking this information on the relationships between states into account—the *ec*-cases—can drastically change the outcomes. Together with the questionable justification linking substitution costs to transition rates, these remarks advocate against using such transition-rate-based methods.

Now, the choice between the remaining solutions will depend on what the researcher is interested in. For instance, when studying the destandardization of family life, the focus may be on changes in the order of successive family life events, in changes in the age at which people experience events such as marriage or the birth of the first child, or changes in durations, such as the time to the birth of the first child after the first union.

If the focus is on changes in sequencing, measures that are highly sensitive to sequencing should be preferred. Good choices are OMstran—OM of transitions—with low weight (i.e. a low w-value) on the state of origin, OMspell—OM of spells—with low expansion cost e and SVRspell—the subsequence vectorial representation metric—with low b spell length weight. One of the differences between these three measures is the sensitivity to small perturbations. If we are interested in these small differences, such as small spells of unemployment, SVRspell should be used. In contrast, OMstran appears to be less sensitive to this aspect, whereas OMspell shows an intermediary position. Classic OM is definitely not suited to measuring differences in sequencing.

If we are interested in explaining changes in timing, then we need measures that are sensitive to timing. Positionwise measures, such as those of the Hamming family, are the most time sensitive. Using the CHI2- and EUCLID-distances with the number of periods K equal to the sequence length is also a solution. This K-parameter offers the advantage of allowing a smooth relaxation of exact timing alignment. This can also be achieved by expressing the Hamming distance as an OM distance with a high indel value, then progressively lowering the indel value. CHI2 is especially interesting when we want to stress the importance of changes involving rare states.

With regard to duration, the CHI2- and EUCLID-distances with K set as 1 are recommended when the interest is primarily in the distribution over the entire period. If the importance of spell lengths needs to be stressed, then OMspell with a high expansion cost would be better. Indeed, LCS and the classic OM distance should also reflect dissimilarities in spell durations reasonably well. Distances of the Hamming, SVRspell- and OMtrans-families are less suited to focusing on differences in spell durations.

Whereas the choice of a dissimilarity measure is relatively easy when the focus remains limited to a single dimension, the choice becomes more difficult when we want to consider differences in two or three dimensions simultaneously. Measures such as OMstran, OMspell and SVRspell, which can cover a large mix of sensitivities, look interesting in this multifocus context, as they allow control of the trade-off between the various dimensions.

In many applications, we may be interested in specific differences that are attributable to each of the timing, duration and sequencing aspects, rather than needing to consider them simultaneously. It could then be useful to use three different dissimilarity measures: one sensitive to timing, one to duration and one to sequencing. This would allow us to identify distinctions stemming from each aspect by comparing the analysis outcomes that are obtained from each measure. For example, when studying the long-term consequences of professional integration trajectories, we would probably look at differences between the trajectories of those who reach stable professional situations and those who stay more vulnerable. Finding greater differences with sequencing-sensitive measures than with timing or duration-sensitive measures would indicate that the effect of the unemployment policy depends more on the order in the trajectory

than on temporality. Similarly, when studying differences in family formation trajectories across birth cohorts, we should be able to determine whether differences are due primarily to changes in sequencing. This may reflect, for instance, the emergence of new stages, such as 'cohabiting couples'. We could also determine whether changes are due to timing differences, perhaps resulting from the postponement of events such as marriage or childbirth, or due to differences in spell durations, such as duration of marriage or the delay between marriage and the birth of the first child.

Running cluster analyses with different dissimilarity measures should also allow us to determine whether the trajectories are primarily structured by timing, duration or sequencing differences. To achieve this, we compare cluster quality measures such as the average silhouette width of the different partitions that are obtained. In a discrepancy analysis, comparing outcomes that are obtained with different measures may also help to identify which covariates best explain sequencing differences, and which best explain timing and duration differences.

However, it is worth recalling that the different dimensions are not completely independent from each other. Therefore, we may observe only minor differences between outcomes that are obtained with different measures. Nevertheless, the use of multiple measures can provide interesting information about borderline cases. For instance, we could learn that a given trajectory looks more like a type A in terms of sequencing, and more like a type B from a timing point of view.

6. Conclusion

Dissimilarity-based sequence analysis has gained much popularity in life course studies in recent years. Although OM remains the most used dissimilarity measure, many other ways of measuring dissimilarity exist. Thus, the researcher faces the difficult task of choosing a suitable measure for her or his research objectives. Our structured and critical review of current measures, together with our simulation study of the behaviour of many of these variants, are intended to help to make this choice.

This review is original in several respects. First, it is specifically oriented towards the ability of the measures to render timing, duration and sequencing differences, which are important in life course studies. Second, it pays attention to the often overlooked mathematical properties of the dissimilarity measures, showing, for instance, that measures that can potentially violate the triangle inequality may exhibit unexpected behaviour. Third, the review covers a unique list of measures.

In this study, we also proposed new distance measures and original strategies to set the costs in OM. OM between sequences of spells has proven to be valuable, notably when considering duration and sequencing. Our reformulation of the OM of the sequences of transitions that was introduced by Biemann (2011) also gave good results in the simulations, covering a wide range of sensitivities depending on the parameters. The strategies that we proposed to set the costs in OM include data-driven indel costs based on state frequencies, state property-based costs by using the Gower distance and common future-based substitution costs.

A few aspects that were not addressed in this study deserve mention here. For example, we did not study the ability of the dissimilarity measures to cope with sequences of unequal length, or to cope with missing elements in the sequences. Further, we did not consider normalized distances. A finer comprehension of how dissimilarity measures behave in these situations is still needed, and we plan to run such a study using a simulation design that is similar to that described here.

To conclude, note that the entire set of dissimilarity measures that were studied with simulated sequences were implemented in the `TraMineR` R package.

Acknowledgements

This publication results from research work that was conducted within the framework of the Swiss National Centre of Competence in Research LIVES Overcoming Vulnerability: Life Course Perspectives (IP14), which is financed by the Swiss National Science Foundation. The authors are grateful to the Swiss National Science Foundation for its financial support. The authors also thank the referees for their constructive comments.

References

Abbott, A. (1983) Sequences of social events: concepts and methods for the analysis of order in social processes. *Hist. Meth.*, **16**, 129–147.
Abbott, A. (1990) A primer on sequence methods. *Organizn Sci.*, **1**, 375–392.
Abbott, A. (1992) From causes to events: notes on narrative positivism. *Sociol. Meth. Res.*, **20**, 428–455.
Abbott, A. (2000) Reply to Levine and Wu. *Sociol. Meth. Res.*, **29**, 65–76.
Abbott, A. and Forrest, J. (1986) Optimal matching methods for historical sequences. *J. Interdisc. Hist.*, **16**, 471–494.
Abbott, A. and Hrycak, A. (1990) Measuring resemblance in sequence data: an optimal matching analysis of musician's careers. *Am. J. Sociol.*, **96**, 144–185.
Abbott, A. and Tsay, A. (2000) Sequence analysis and optimal matching methods in sociology, Review and prospect (with discussion). *Sociol. Meth. Res.*, **29**, 3–76.
Aisenbrey, S. and Fasang, A. E. (2010) New life for old ideas: the "second wave" of sequence analysis bringing the "course" back into the life course. *Sociol. Meth. Res.*, **38**, 430–462.
Bergroth, L., Hakonen, H. and Raita, T. (2000) A survey of longest common subsequence algorithms. In *String Processing and Information Retrieval, 2000: Proc. 7th Int. Symp*., pp. 39–48. A Curuna: Institute of Electrical and Electronics Engineers.
Biemann, T. (2011) A transition-oriented approach to optimal matching. *Sociol. Methodol.*, **41**, 195–221.
Billari, F. C., Fürnkranz, J. and Prskawetz, A. (2006) Timing, sequencing, and quantum of life course events: a machine learning approach. *Eur. J. Popln*, **22**, 37–65.
Bras, H., Liefbroer, A. C. and Elzinga, C. H. (2010) Standardization of pathways to adulthood?: an analysis of Dutch cohorts born between 1850 and 1900. *Demography*, **47**, 1013–1034.
Deville, J.-C. and Saporta, G. (1983) Correspondence analysis with an extension towards nominal time series. *J. Econmetr.*, **22**, 169–189.
Dijkstra, W. and Taris, T. (1995) Measuring the agreement between sequences. *Sociol. Meth. Res.*, **24**, 214–231.
van Driel, K. and Oosterveld, P. (2001) Nonoptimal alignment: a comment on "Measuring the agreement between sequences" by Dijkstra and Taris. *Sociol. Meth. Res.*, **29**, 524–531.
Elzinga, C. H. (2003) Sequence similarity: a non-aligning technique. *Sociol. Meth. Res.*, **31**, 214–231.
Elzinga, C. H. (2005) Combinatorial representations of token sequences. *J. Classificn*, **22**, 87–118.
Elzinga, C. H. (2007) Sequence analysis: metric representations of categorical time series. *Manuscript*. Department of Social Science Research Methods, Vrije Universiteit, Amsterdam.
Elzinga, C. H., Rahmann, S. and Wang, H. (2008) Algorithms for subsequence combinatorics. *Theor. Comput. Sci.*, **409**, 394–404.
Elzinga, C. H. and Studer, M. (2015) Spell sequences, state proximities and distance metrics. *Sociol. Meth. Res.*, **44**, 3–47.
Gabadinho, A. and Ritschard, G. (2013) Searching for typical life trajectories applied to childbirth histories. In *Gendered Life Courses—between Individualization and Standardization; a European Approach applied to Switzerland* (eds R. Levy and E. Widmer), pp. 287–312. Vienna: LIT.
Gauthier, J.-A., Widmer, E. D., Bucher, P. and Notredame, C. (2009) How much does it cost?: optimization of costs in sequence analysis of social science data. *Sociol. Meth. Res.*, **38**, 197–231.
Gower, J. C. (1971) A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 857–874.
Grelet, Y. (2002) Des typologies de parcours: méthodes et usages. *Notes de Travail Génération 92*. Céreq, Paris.
Halpin, B. (2010) Optimal matching analysis and life-course data: the importance of duration. *Sociol. Meth. Res.*, **38**, 365–388.
Halpin, B. (2014) Three narratives of sequence analysis. In *Life Course Research and Social Policies Advances in Sequence Analysis: Theory, Method, Applications* (eds P. Blanchard, F. Bühlmann and J.-A. Gauthier), vol. 2, pp. 75–103. Berlin: Springer.
Hamming, R. W. (1950) Error detecting and error correcting codes. *Bell Syst. Tech. J.*, **26**, 147–160.
Henikoff, S. and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proc Natn. Acad. Sci. USA*, **89**, 10915–10919.
Hogan, D. P. (1978) The variable order of events in the life course. *Am. Sociol. Rev.*, **43**, 573–586.

Hollister, M. (2009) Is optimal matching suboptimal? *Sociol. Meth. Res.*, **38**, 235–264.
Kruskal, J. B. (1983) An overview of sequence comparison: time warps, string edits, and macromolecules. *SIAM Rev.*, **25**, 201–237.
Laub, J. and Müller, K.-R. (2004) Feature discovery in non-metric pairwise data. *J. Mach. Learn. Res.*, **5**, 801–818.
Lesnard, L. (2010) Setting cost in optimal matching to uncover contemporaneous sociotemporal patterns. *Sociol. Meth. Res.*, **38**, 389–419.
Levenshtein, V. (1966) Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.*, **10**, 707–710.
Levine, J. (2000) But what have you done for us lately. *Sociol. Meth. Res.*, **29**, 35–40.
Massoni, S., Olteanu, M. and Rousset, P. (2009) Career-path analysis using optimal matching and self-organizing maps. In *Advances in Self-organizing Maps: 7th Int. Wrkshp, St Augustine, June 8th–10th* (eds J. C. Príncipe and R. Mikkulainen), pp. 154–162. Berlin: Springer.
Needleman, S. and Wunsch, C. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Molec. Biol.*, **48**, 443–453.
Notredame, C., Bucher, P., Gauthier, J.-A. and Widmer, E. D. (2006) T-COFFEE/SALTT: user guide and reference manual. *Technical Report*. Centre National de la Recherche Scientifique, Marseille. (Available from `http://www.tcoffee.org/saltt/`.)
Pollock, G. (2007) Holistic trajectories: a study of combined employment, housing and family careers by using multiple-sequence analysis. *J. R. Statist. Soc.* A, **170**, 167–183.
Robette, N. and Bry, X. (2012) Harpoon or bait?: a comparison of various metrics in fishing for sequence patterns. *Bull. Sociol. Methodol.*, **116**, 5–24.
Rohwer, G. and Poetter, U. (2005) *TDA User's Manual*. Bochum: Ruhr Universität Bochum.
Rousset, P., Giret, J.-F. and Grelet, Y. (2012) Typologies de parcours et dynamique longitudinale. *Bull. Sociol. Methodol.*, **114**, 5–34.
Schumacher, R., Matthijs, K. and Moreels, S. (2012) Migration and reproduction in an urbanizing context: a sequence analysis of family life courses in 19th century Antwerp and Geneva. *Working Papers 17*. Wetenschappelijke Onderzoeksgemeenschap Historical Demography, Leuven.
Settersten, R. A. J. and Mayer, K. U. (1997) The measurement of age, age structuring, and the life course. *A. Rev. Sociol.*, **23**, 233–261.
Stovel, K. (2001) Local sequential patterns: the structure of lynching in the deep south, 1882-1930. *Socl Forces*, **79**, 843–880.
Stovel, K. and Bolan, M. (2004) Residential trajectories: using optimal alignment to reveal the structure of residential mobility. *Sociol. Meth Res.*, **32**, 559–598.
Studer, M. (2012) Étude des inégalités de genre en début de carrière académique à l'aide de méthodes innovatrices d'analyse de données séquentielles. In *Collection des Thèses*, vol. SES-777. Facultè des Sciences Économiques et Sociales, Université de Genève, Genève.
Studer, M. (2013) WeightedCluster library manual: a practical guide to creating typologies of trajectories in the social sciences with R. *Working Paper 24*. Swiss National Center of Competence in Research LIVES, Geneva.
Studer, M. and Ritschard, G. (2014) A comparative review of sequence dissimilarity measures. *Working Paper 33*. Swiss National Center of Competence in Research LIVES, Geneva.
Studer, M., Ritschard, G., Gabadinho, A. and Müller, N. S. (2011) Discrepancy analysis of state sequences. *Sociol. Meth. Res.*, **40**, 471–510.
Todorov, V. and Filzmoser, P. (2009) An object-oriented framework for robust multivariate analysis. *J. Statist. Softwr.*, **32**, 1–47.
Widmer, E. D., Levy, R., Pollien, A., Hammer, R. and Gauthier, J.-A. (2003) Between standardisation, individualisation and gendering: an analysis of personal life courses in Switzerland. *Swiss J. Sociol.*, **29**, 35–65.
Widmer, E. and Ritschard, G. (2009) The de-standardization of the life course: are men and women equal? *Adv. Lif. Course Res.*, **14**, 28–39.
Wilson, C. (2006) Reliability of sequence-alignment analysis of social processes: Monte Carlo tests of ClustalG software. *Environ. Planng* A, **38**, 187–204.
Wu, L. L. (2000) Some comments on 'Sequence analysis and optimal matching methods in sociology: review and prospect'. *Sociol. Meth. Res.*, **29**, 41–64.
Yujian, L. and Bo, L. (2007) A normalized Levenshtein distance metric. *IEEE Trans. Pattn Anal. Mach. Intell.*, **29**, 1091–1095.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Detailed simulations results';
'Appendix: Chronograms of groups of simulated sequences'.