L I V E S
W O R K I N G
P A P E R
2 0 1 4 / 33

TITLE

A comparative review of sequence dissimilarity measures

Research paper

Authors

Matthias Studer

Gilbert Ritschard

http://dx.doi.org/10.12682/lives.2296-1658.2014.33
ISSN 2296-1658

FNSNF
Swiss National Science Foundation

The National Centres of Competence in Research (NCCR) are a research instrument of the Swiss National Science Foundation.

LIVES
Swiss National Centre of Competence in Research

Authors

Studer, M. (1)

Ritschard, G. (2)

Abstract

This is a comparative study of the multiple ways of measuring dissimilarities between state sequences. For sequences describing life courses, such as family life trajectories or professional careers, the important differences between the sequences essentially concern the sequencing (the order in which successive states appear), the timing, and the duration of the spells in the successive states. Even if some distance measures underperform, it has been shown that there is no universally optimal distance index and that the choice of a measure depends on which aspect we want to focus on. This study also introduces novel ways of measuring dissimilarities that overcome the flaws in existing measures.

Keywords

State sequences | Dissimilarity | Distance | Optimal matching | Sequencing | Timing | Duration | Spells

Author's affiliation

(1) (2) LIVES & IDEMO, University of Geneva

Correspondence to

matthias.studer@unige.ch

FNSNF
SWISS NATIONAL SCIENCE FOUNDATION

The National Centres of Competence in Research (NCCR) are a research instrument of the Swiss National Science Foundation.

LIVES
Swiss National Centre of Competence in Research

A comparative review of sequence dissimilarity measures

Matthias Studer Gilbert Ritschard

LIVES and IDEMO, University of Geneva

Abstract

This is a comparative study of the multiple ways of measuring dissimilarities between state sequences. For sequences describing life courses, such as family life trajectories or professional careers, the important differences between the sequences essentially concern the sequencing (the order in which successive states appear), the timing, and the duration of the spells in the successive states. Even if some distance measures underperform, it has been shown that there is no universally optimal distance index and that the choice of a measure depends on which aspect we want to focus on. This study also introduces novel ways of measuring dissimilarities that overcome the flaws in existing measures.

Keywords: state sequences, dissimilarity, distance, optimal matching, sequencing, timing, duration, spells.

Author's note: Please address correspondence to Matthias Studer, Institute for Demographic and Life Course Studies, University of Geneva, Bvd Pont D'Arve 40, 1211 Geneva 4, Switzerland, Matthias.Studer@unige.ch.

1 Introduction

Since Andrew Abbott (1983) stressed the relevance of sequence methods to the social sciences, sequence analysis, and particularly the so-called optimal matching analysis (Abbott and Forrest, 1986; Abbott and Hrycak, 1990), has become popular. 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; Schumacher et al., 2012)). The strength of the sequence approach is the holistic view it provides by dealing with whole trajectories. This allows us to determine trajectory patterns that account for all states of interest experienced during the considered period. This contrasts, for instance, with survival or event history analysis, which, by focusing on the hazard of—or time to—a specific event, does not give an overall view of how the trajectories are organised.

The so-called optimal matching analysis measures pairwise dissimilarities between sequences, and then identifies 'types' of patterns by clustering the sequences based on these dissimilarities. Beneath clustering analysis, other dissimilarity-based methods have also proven useful when investigating sequence data. For instance, Abbott (1983) mentions multidimensional scaling, Massoni et al. (2009) use self-organising maps, Studer et al. (2011) show how to run ANOVA-like analyses and to grow regression trees on sequence data, and Gabadinho and Ritschard (2013) search for non-redundant typical patterns with the densest neighbourhoods.

Despite often being referred to as 'optimal matching analysis', so named by Abbott and Forrest (1986) after the edit distance they used, dissimilarity-based analysis is in no way restricted to optimal matching (OM) distances. The methods also work with other ways of measuring dissimilarity, and, as we will see, many different distances have been proposed. For example, there are Chi-square distances adapted for sequence data that essentially measure differences in state distributions, distances based on counts of common attributes (e.g. matching states or subsequences), and multiple variants of editing dissimilarity measures, such as OM, which evaluates differences according to the cost of 'editing' one sequence into the other.

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, the choice of a dissimilarity measure deserves special attention, and it is important that we understand what we want to account for before quantitatively evaluating the difference between two sequences. This study aims to contribute to this understanding by identifying the various aspects (e.g. constituting 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 organised 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. Lastly, we also propose novel dissimilarity measures, either because there is a lack of existing measures that can consistently account for some aspect, or as variants of existing measures to fix some of their weaknesses. We thus introduce a new edit measure—OM between sequences of spells—to consistently account for differences in the time spent in the distinct successive states (DSS). We also propose an original solution to set data-driven indel costs. Among the improvements to existing measures, we suggest using the Gower distance to

determine the substitution costs for a mix of quantitative and qualitative attributes of the states. In addition, we propose defining substitution costs as a Chi-square distance, stressing the similarity between states that share the same future. This is an alternative to the questionable derivation of substitution costs based on transition rates. Lastly, we propose solutions that drastically reduce the number of control parameters in the OM of the sequences of transitions introduced by Biemann (2011).

The remainder of this paper is organised as follows. We first set the framework by specifying the kinds of sequences we consider, the different aspects we may want the dissimilarity measures to reflect, and by recalling distance-related notions of interest. We then present the dissimilarity measures we wish to consider, as well as their theoretical properties. In the following section, we examine the behaviour of the measures using artificially generated data, and 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 presented in this paper are available in the latest version of the TraMineR R library. See Appendix A for a brief explanation of how to compute these measures.

2 Sequences and distances

Definitions and notation. We consider categorical sequences, defined as an ordered list of successive elements chosen from a finite alphabet, Σ. In demography or sociology, and more generally, for sequences describing life trajectories, the elements in the sequences are 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 would conclude that the individual worked full time for ten years.

The natural way to encode a state sequence is to list the successive elements. For example, the trajectory of someone who is 2 years 'Junior Manager' (JM) and then 3 years 'Senior Manager' (SM) is represented as JM-JM-SM-SM-SM. We can also encode the sequence in a more compact way as $\text{JM}^2\text{-SM}^3$. In other words, we simply list the DSS and add a duration stamp—in this case, as the subscript—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 the duration of spells in the same state. In the remainder of this paper, we shall use the term *spell* to refer to the whole spell spent in the same state.

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, p. 207) distinguishes four different ways—in fact transformation operations—in which sequences may 'differ': substitutions, deletions and insertions (indels), compression 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 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, the interest when comparing sequences is not in the transformation of one sequence to another, but more directly in how the sequences differ in socially meaningful aspects. In line with the distinctions made by Settersten and Mayer (1997) and Billari et al. (2006), we identify the following important aspects.

The first basic aspect of interest when comparing the trajectories of two individuals is the *list of distinct states each experiences*. This is what Dijkstra and Taris (1995) and Elzinga (2003) implicitly refer to when stating that two sequences with no common state are maximally dissimilar.[1] The notion of experienced states is also related to the quantum 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 impact of total exposure times. For instance, we may want to examine the effect of the total amount of time spent as unemployed on the person's health status at retirement. However, differences in the presence/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 events occur, such as the start of a spell (Settersten and Mayer, 1997) in a given state, is a sociologically important aspect. Life course literature often stresses, for instance, the role of age norms in the construction of life trajectories (Widmer et al., 2003). Moreover, the social reality reflected by a state often depends on its position in the trajectory. For example, Rousset et al. (2011) observe that the impact of unstable employment on the professional insertion trajectory increases with age. In addition, Lesnard (2010) claims, in his study on the way couples use time, that differences between 'no partner working' and 'only one partner working' reflect a very different reality when observed during the day or night. *Spell duration*, the consecutive time spent in the same state, is another way to account for time (Settersten and Mayer, 1997).[2] 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 getting married, or the duration of a joblessness episode, are important aspects within people's life courses. Spell duration is different to the information provided by the state distribution in that it gives the consecutive exposure time, rather than the total—not necessarily consecutive—exposure time. Unlike the state distribution, the spell duration allows us to, for example, distinguish between long-term joblessness 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 emphasised by, for example, Hogan (1978). Experiencing childbirth before or after marriage reflects different ways of life. Abbott (1990) identifies sequencing as the key concept in sequence analysis, and Billari et al. (2006) emphasise its importance in conjunction with timing

[1] Note that this claim would not hold if some states can be considered to be more similar than others.

[2] Settersten and Mayer (1997) even consider the more general concept of *spacing* to refer to the time between any two events or transitions.

and quantum for demographic life course analysis.

To summarise, we can distinguish at least five socially sound aspects on which two sequences may differ, namely:

1. *Experienced states*: the distinct elements of the alphabet present in the sequence
2. *Distribution*: the within-sequence state distribution (total time)
3. *Timing*: the age or date at which each state appears
4. *Duration*: the spell lengths in the DSS
5. *Sequencing*: the order of the DSS.

In the discussion that follows, we denote occupational sequences as comprising education (E), unemployment (U), full-time work (F), and part-time work (P).

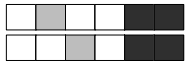
The five aforementioned aspects are not independent of each other. For example, by changing the sequencing, we also change timing, and changing consecutive times in the states implies changes in the within-sequence distribution, and possibly in the sequencing as well. Likewise, modifying the within-sequence distribution by changing the distinct present states affects the sequencing and durations. From the reverse point of view, two sequences that are similar on one aspect may be quite different on another. For example, the sequences EUUUUF and EUUFUU are similar in terms of the involved states and the within-sequence distribution. However, they differ in the consecutive time spent in the DSS and in the sequencing and timing of F and one U. Table 1 summarises the dependence relationships between the five facets of sequence differences. The list of 'States' found in the sequences automatically follows from the non-zero frequencies in the 'Distribution' of the states, as well as from the sequencing of states (DSS). In addition, 'Durations', the consecutive time spent in the states, and 'Timing' only make sense when associated with the relevant sequence of DSS. Specifying the sequencing—the DSS—together with either the timing or the duration of the DSS entirely defines the sequence.

Table 2 gives examples of pairs of sequences that look similar from one point of view, but differ from another. For instance, for a given distribution, we still have degrees of

Table 1. Dependence relationships between sequence characteristics ('States' stands for experienced states, 'Distribution' for the state distribution within the sequence, 'Sequencing' for the order of the DSS, 'Durations' for the consecutive times spent in the distinct states, 'Timing' for the position of the states in the sequence, and 'Sequence' for a wholly characterised sequence)

Distribution	→	States
Sequencing	→	States
Durations	→	Distribution
Sequencing + Durations	→	Sequence (States, Distribution, Timing)
Sequencing + Timing	→	Sequence (States, Distribution, Durations)

Table 2. Examples of pairs of sequences similar from one point of view, but different from another. The '=' sign indicates that when the sequences have the same row characteristic, they cannot differ in the column characteristic

	Different in			
Same	Distribution	Spell Durations	Timing	Sequencing
States				
States + Distribution	=			
Sequencing				=
Sequencing + Distribution	=			=

freedom on spell durations, timings, and sequencings. The professional career of someone having experienced 12 months of joblessness may reflect greater instability when the 12 months of unemployment are split into several short unemployment periods than when the 12 months are contiguous. In that case, measuring only the differences in distributions would be insufficient. Likewise, focusing only on the state order—the sequencing—leaves room for different durations and timings, which we may actually want to consider.

In addition to the listed aspects, we may be interested in higher-order characteristics, such as the transitions between states. We can then compare sequences of higher-order characteristics using the same criteria, when applicable. For instance, spell durations are not applicable to transitions. As an example, the transition-based distance proposed by Biemann (2011) attempts to evaluate the dissimilarity between sequences of transitions rather than directly from the state sequences from which they are derived.

Measuring dissimilarity. A measure of dissimilarity is a quantitative evaluation of the level of mismatch between two sequences. This level of mismatch gives an idea of the (lack of) resemblance between the sequences and, therefore, is useful for comparing sequences.

Dissimilarity measures can be roughly classified into three classes: i) Distances between distributions; ii) Measures based on the count of common attributes between sequences; and iii) Edit distances, which measure the cost of the operations necessary to transform one sequence into the other. By specifying the distributions or the attributes used to count the (mis)matches, the two former classes more or less explicitly specify the aspect they focus on. For example, by comparing state distribution within sequences, the focus is on durations, while counting position-wise mismatches focuses on the exact timing equality of matching states. In contrast, the focus is less evident for edit distances, which are based on elementary transformation operations that may impact the timing, duration, or order of the states.

One additional concern is whether the measure should account for the involved ele-

ments or be based uniquely on the count of matches and mismatches. For example, should we consider FFUUU as more different to EEUUU than to PPUUU? The simple count of position-wise mismatches is insensitive to which tokens do not match, and would lead to the same dissimilarity for any pair of the sequences FFUUU, EEUUU, and PPUUU. On the other hand, accounting for distances between states renders dissimilarities sensitive to which of the elements do not match. Considering P and F as more similar than E and F, FFUUU would be closer to PPUUU than to EEUUU.[3]

Distance-related mathematical notions. To correctly interpret the outcome of dissimilarity-based methods, for instance, to be sure a cluster solution minimises the within-discrepancy or to ensure the non-negative definiteness of the matrix to be diagonalised in multidimensional scaling, dissimilarity measures usually need to be distances, or even Euclidean distances. Therefore, it is worthwhile recalling these mathematical properties here.

A dissimilarity measure, $d(x, y)$, between two sequences, x and y, is a *distance* if and only if it fulfills the following conditions: $d(x, y) \geq 0$, $d(x, y) = 0$ if and only if $x = y$, $d(x, y) = d(y, x)$ for any x, y (symmetry), and $d(x, y) \leq d(x, z) + d(z, y)$ for any x, y, z (triangle inequality). Discrepancy analysis and most clustering algorithms require symmetrical dissimilarities. 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 $d(x, y)$ because of a third sequence, z. In this case, the actual dissimilarity would depend on the other sequences present in the data set (Elzinga and Studer, 2013).

In a Euclidean space with n real coordinates, the Euclidean distance between two points, x and y, is defined as $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$, and corresponds in two dimensions to the distance that would be measured with a ruler.

As defined above, sequences do not, in their original form, have real coordinates, so we cannot directly apply Euclidean distances to them. Two strategies are then possible.

The first one characterises each sequence using a series of numerical attributes and then computes the distance between those vectors of attributes. In such approaches, known as kernel methods (Lodhi et al., 2002), the attributes are often only theoretically defined, and the distance is computed in an efficient way without the need to actually determine the individual values of the attributes. However, by definition, we know the distance is Euclidean. A distance between any kind of objects is Euclidean if and only if we can map the objects onto real coordinates such that the Euclidean distance between any of the resulting numerical points reproduces exactly the distance between the objects they represent (Gower, 1982). The Euclidean property is of particular interest in multidimensional scaling.[4]

[3]The latter approach has been used for multichannel sequences, which can be viewed as single-channel sequences made up of states resulting from the combination of states from the different channels. For example 'Married and working full time' combines a state from the family life with a state from the occupational channel. Some authors (e.g., Pollock, 2007) define state-dependent multichannel distances using dissimilarities between combined states, themselves derived from the state dissimilarities within each single channel.

[4]For non-Euclidean distances, multidimensional scaling produces complex coordinates associated with negative eigenvalues. These (usually ignored) complex coordinates reflect the distortion incurred by embedding sequences into a Euclidean vector space. Therefore, it may be worth studying them (see for

The second strategy is to define an ad hoc dissimilarity measure—e.g. using optimal matching—that accounts for specific aspects of sequence differences. Such dissimilarity measures may be more in phase with the domain of the analysis, but they might not be Euclidean, or could even violate the symmetry and/or triangle inequality requirements.

3 Overview of dissimilarity measures

In this section, we briefly describe the main measures, stressing their aims and properties, and how we can expect them to behave with respect to differences in timing, duration, and sequencing. We start by addressing distances between within-sequence state distributions, then consider measures based on the count of common attributes, and, lastly, discuss OM and other related edit distances.

3.1 Distances between probability distributions

Distances between state distributions. The first approach used to measure 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, on the time spent in each state within the sequences. For example, if the alphabet is {E,F,P,U}, the state distribution of the sequence EFEFFUU is $(2/7, 3/7, 0, 2/7)$. Given these distribution vectors, the dissimilarity between sequences is measured by the distance between the vectors using either the Euclidean distance or the Chi-square distance. The former accounts for the absolute differences in the proportion of time spent in the states. The squared Chi-square 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, gives more importance to a rare state than to a frequent state. Letting $p_{j|x}$ be the proportion of time spent in state j in sequence x, and p_j the overall proportion of time spent in state j, the squared Chi-square distance reads as follows:

$$d^2_{chi}(x,y) = \sum_{j=1}^{|\Sigma|} \frac{(p_{j|x} - p_{j|y})^2}{p_j}. \tag{1}$$

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. Denoting the subsequence of x over period k as x_k, and the overall proportion of time in state j in the kth interval as $p_{j|k}$, the period-dependent Chi-square distance becomes

$$d^2_{chi,K}(x,y) = \sum_{k=1}^{K} \sum_{j=1}^{|\Sigma|} \frac{(p_{j|x_k} - p_{j|y_k})^2}{p_{j|k}}. \tag{2}$$

In other words, the distance is the sum of the Chi-square distances for each period. At the limit, when K is equal to the length, ℓ, of the sequences, the value of each $p_{j|x_k}$ can

example Laub and Müller, 2004).

only be 0 or 1, and 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.

Distance based on conditional distributions of subsequent states. A related measure, defined as the sum of the position-dependent distances computed at successive positions, was proposed by Rousset et al. (2012) to measure the dissimilarity between sequences describing professional integration trajectories. The aim of this measure is to stress the similarity of the sequences 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 Chi-square distance between the vectors of the—weighted and normalised—transition rates from the state observed at t to the states observed at the subsequent positions, $t+1, t+2, ...$

Formally, the distance between two sequences, x and y, is the sum, $\sum_{t=1}^{\ell} d_t(x_t, y_t)$, of the distances between the pairs of states x_t and y_t observed at the successive positions, t. Let a_t be state $a \in \Sigma$ at t (what the authors call 'situation a_t'), and $\mathbf{p}_{a_t}$ be the vector collecting the $\ell|\Sigma|$ weighted and normalised transition rates, $\tilde{p}(b_{t'}|a_t)$, from a_t to $b_{t'}$, for all $t' = 1, ..., \ell$ and $b \in \Sigma$. Each transition rate is weighted by a decreasing function of the time interval, $t' - t$ (e.g. $1/(t' - t + 1)$), and the weighted rates are normalised to sum to one over b and t', such that $\sum_{t'=1}^{\ell} \sum_{b \in \Sigma} \tilde{p}(b_{t'}|a_t) = 1$. The distance $d_t(x_t, y_t)$ at t is the Chi-square distance between the normalised-weighted vectors $\tilde{\mathbf{p}}_{x_t}$ and $\tilde{\mathbf{p}}_{y_t}$ corresponding to the states in x and y at position t. The role of the weight is to give more importance to the near future than the far future when evaluating the distance between the states at position t. For example, the probability of being a manager one year after high school when aged 25 will get a higher weight than the probability of being a manager 10 years later. The normalisation is necessary for the vectors $\mathbf{p}_{x_t}$ to look like probability distributions. This is in spite of the weights and the values being constrained to be null when, as in Rousset et al. (2012), only transitions to future states are considered and all transition rates from t to $t' < t$ are assumed to be zero.

Since the distance defined by Rousset et al. is the sum of position-wise distances, it should, like the position-wise Euclidean and Chi-squared distances, be sensitive to non-matching timings and differences in sequencing. However, we can expect that the introduced link to the future will smooth this sensitivity somewhat.

The Chi-square 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 defined as the sum of the position-wise distances between states only apply to pairs of sequences of same length.

3.2 Distances based on counts of common attributes

The number of common attributes, $A(x, y)$, between sequence x and y measures proximity. In other words, a larger value of $A(x, y)$ implies that the sequences are closer together. A common way of turning such a proximity into a dissimilarity measure is as follows:

$$d_A(x, y) \quad = \quad A(x, x) + A(y, y) - 2A(x, y). \tag{3}$$

Simple Hamming distance. Hamming (1950) proposed measuring the dissimilarity between two sequences by the number of positions with non-matching states. The Hamming distance, $d_H(x, y)$, can be expressed in the form of Eq. 3 by setting $A(x, y)$ as half the number of mismatches between x and y.[5] Since the Hamming distance proceeds by a position-wise 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.

The length of the longest common prefix, suffix, and subsequence. Distances derived from the length of the longest common prefix or suffix are further typical examples of squared Euclidean distances based on the count of common attributes (Elzinga, 2007). The length of the longest common subsequence (LCS; e.g. see (Bergroth et al., 2000)) 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. The derived d_{LCS} distance is obtained by setting $A(x, y)$ to the number of elements matching in this way. 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 position-wise 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 the 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.

Number of matching subsequences. Elzinga (2003, 2005) introduced a dissimilarity measure, d_{NMS}, based on the number of matching subsequences. 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. Letting $\mathrm{emb}_z(u)$ be the number of times subsequence u is embedded in sequence z, the number of matching subsequences, $A_{\mathrm{NMS}}(x, y)$, between x and y is

$$A_{\mathrm{NMS}}(x, y) = \sum_{u \in S(x,y)} \mathrm{emb}_x(u)\, \mathrm{emb}_y(u),$$

where $S(x, y)$ denotes the set of distinct common subsequences. For each embedding of u in x, we count how many times it can be matched with a different embedding of u in y.

The measure has been extended in different ways, for instance, by weighting the length of the subsequence u and/or considering the minimal shared time in the pair of embedded subsequences (Liefbroer and Elzinga, 2012). The last and most general evolution of this measure is the *subsequence vector representation-based metric* (SVR), introduced by Elzinga and Studer (2013). The SVR allows us to weight each matched subsequence u by its length ℓ_u, or a transformation ℓ_u^a of it, as well as to account for the duration of each embedding. Each subsequence, u, is weighted by the product of the sums of the durations of the embeddings, u, in each of the two sequences. For example, the sum of the durations of the subsequence E-U in the sequence E^5-U^4-F^6 is $5+4=9$, but is $(7+2)+(7+1)=17$ for the sequence E^7-F^2-U^2-F^3-U^1. The subsequence E-U will therefore receive a weight

[5] The Hamming distance also corresponds to the Gower distance with equally weighted states and positions, as considered by Wilson (2006).

of $9 \cdot 17 = 153$ when comparing the two sequences. Elzinga and Studer (2013) propose fine-tuning the weights with a parameter b using a transformation, t^b, instead of the spell duration, t. A higher b should give comparatively more weight to longer spells. In addition to these weighting mechanisms, the SVR can also account for state proximities. This is the version we shall consider as the d_{SVRspell} distance in Section 5.

Both the original distance, d_{NMS}, and d_{SVRspell}, accounting for the duration, are Euclidean distances. By construction, we should expect both distances to be very sensitive to differences in sequencing. The measures should also be sensitive to differences in durations. The original version, because the duration extension increases the number of embeddings of concerned subsequences. The second form does so by explicitly considering the duration of spells. Computing d_{NMS} between the sequences of DSS—which is equivalent to d_{SVRspell} with $b = 0$—should, in contrast, be insensitive to differences in timing and durations.

With the exception of the SVR metric, the above distances, based on the count of (mis)matching features or differences between distributions, do not account for the level of dissimilarity between the specific involved states. In contrast, as we show below, the edit distances naturally allow for state-dependent solutions that can be controlled using the individual costs of basic edit operations.

3.3 Optimal matching

Since Andrew Abbott (Abbott and Forrest, 1986; Abbott and Hrycak, 1990) popularised optimal matching 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—Kruskal (1983) details nine different application fields—that use similar edit approaches, such as the Levenshtein distance (Levenshtein, 1966; Yujian and Bo, 2007) in computer science and sequence alignment in bioinformatics.

3.3.1 OM 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—either inserts or deletes—of tokens or substitutions between tokens. Each operation is assigned a cost, which may vary with the involved states. The main criticism of OM is the lack of sociological meaning of these operations and their costs (Abbott and Tsay, 2000; Abbott, 2000; Levine, 2000; Wu, 2000; Aisenbrey and Fasang, 2010; Lesnard, 2010).

Using the formalism of Yujian and Bo (2007), the OM distance can formally be defined as follows. Let Σ be the alphabet—the list of elements that can appear in the sequences—and λ be the null character. The following are valid transformations (basic edit operations): substituting a with b ($a \to b$), deleting a ($a \to \lambda$), and inserting a ($\lambda \to a$), with $a, b \in \Sigma$ and $a \neq b$. Let $T^j_{x,y} = T^j_1 \dots T^j_{\ell_j}$ be a sequence of ℓ_j transformations that turns sequence x into y and $\gamma(T^j_i)$ be the cost of each elementary transformation,

T_i^j. The OM dissimilarity is then

$$d_{OM}(x, y) = \min_j \sum_{i=1}^{\ell_j} \gamma(T_i^j). \tag{4}$$

Let $\mathbf{\Gamma}$ be the $(|\Sigma| + 1) \times (|\Sigma| + 1)$ matrix of the substitution and indel costs. For consistency, the costs, $\mathbf{\Gamma}$, should define a metric between the admissible states. In other words, the costs should be symmetric, should fulfil the triangle inequality, and be zero 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. For example, in Table 3, the value $\gamma(a \to c) = 10$ is inconsistent because we can achieve the same substitution by deleting a and then inserting c for a total cost of only 2. Moreover, existing algorithms, such as the dynamic programming algorithm of Needleman and Wunsch (1970), used to compute the OM distance, all assume the costs satisfy the metric properties. Therefore, they could well return a solution that does not reflect the minimum cost if these properties are violated.

Unless the cost matrix is inconsistent, the OM dissimilarity is necessarily a metric. The OM distance is zero if and only if the two sequences are identical, since every change has a non-zero cost and the cost of no changes is zero. The symmetry of the OM dissimilarity results from the symmetry of the indel and substitution costs. The triangle inequality follows from the definition of the OM distance as a minimum cost, meaning there is no other sequence of transformations that will turn one sequence into the other at a lower cost. As the solution to a minimisation process, the OM distance cannot be expressed as a kernel and, therefore, is not Euclidean (Elzinga, 2007). As a result, OM distances cannot usually be equated to distances between sequence representations on real coordinates.

The parameterisation of OM using the costs of the elementary operations makes it a very flexible dissimilarity measure that can cope with many different situations. For instance, setting arbitrarily high indel costs renders the dissimilarity extremely time sensitive. In contrast, low indel costs—with respect to substitution costs—downweight the importance of time shifts in sequence comparisons. Costs also allow for state-dependent dissimilarities between sequences. On the other hand, the lack of definitively sound sociological rules for setting the costs is one of the more criticised aspects of OM (Levine, 2000; Wu, 2000). Furthermore, the high number of indel and substitution costs may be seen as an overparameterisation (Wu, 2000).

To understand the impact of the costs, we investigate two special cases of OM.

Table 3. A $\mathbf{\Gamma}$ matrix violating the triangle inequality

	a	b	c	λ
a	0			
b	2	0		
c	10	2	0	
λ	1	1	1	0

Generalised Hamming distance. The simple Hamming distance is equivalent to OM with all substitution costs equal to one and no indels. The generalisation allows for state-dependent substitution costs and is therefore OM without indels. The generalised Hamming distance is the weighted sum of position-wise mismatches between two sequences.

Table 4 illustrates the Hamming distance. Simple Hamming corresponds to constant costs of 1, in which case the distance between x and y is 3. If we consider b and c to be twice as close as a and c, we can set the cost of the substitution $b \to c$, and $c \to b$ by symmetry, as 0.5. The generalised Hamming distance is then 2 instead of 3.

Table 4. Substitution costs: Example

Sequence x	a	a	c	b	c	
Sequence y	a	c	b	b	b	
Constant Cost	0	1	1	0	1	Dist = 3
State-dependent Costs	0	1	0.5	0	0.5	Dist = 2

Levenshtein II. In contrast to the Hamming distance, which does not allow for indels, we can discard substitutions and measure the distance by counting the number of indels necessary to transform one sequence into another. The resulting measure is known as the Levenshtein II distance. This distance measure is equivalent to the LCS distance. The total number of indels required is the sum of the elements in each sequence that are not involved in the LCS, $(|x| - \ell_{\text{LCS}}(x, y)) + (|y| - \ell_{\text{LCS}}(x, y))$, where $\ell_{\text{LCS}}(x, y)$ is the length of the LCS. Indels allow for a time warp in sequence comparisons. A time warp should permit us to find similarities between sequence patterns that differ only in timing and/or state durations, for example, between EFEFFUU and EEFEFFU.

Table 5 illustrates how to align two sequences using only indels. The row 'Operations' indicates the required delete, d, and insert, i, operations on x necessary to turn this sequence into y. We first delete one of the leading states, a, and the c occurring after the b. Then, we append two bs at the end. This is a total of four operations. The LCS is a-c-b and is of length 3. Since both sequences have length 5, we have $d_{\text{LCS}}(x, y) = 5{+}5{-}2{\cdot}3 = 4$. In other words, d_{LCS} is, as expected, equal to the Levenshtein II distance.

Table 5. Distance of Levenshtein II: Example

Sequence x	a	a	c	b	c	-	-	
Operations	d				d	i	i	
Sequence y	-	a	c	b	-	b	b	
Cost	1	0	0	0	1	1	1	Dist = 4

Instead of a single indel cost, it is possible to make the costs state-dependent in the Levenshtein II distance and, more generally, in OM distance. For example, in her study

of lynching in the Deep South, in which she uses OM, Stovel (2001) sets the indel costs as a linear function of the number of yearly lynchings, and gives the lower cost of 1 to the most common '0 lynching' state. Hollister (2009) also suggests assigning a relatively low cost to inserting or deleting what could be considered a normal or default state.

General OM example. Table 6 shows an example of how to compute an OM distance. Using the state-dependent substitution costs from Table 4, with a fixed indel cost of 0.5, the depicted operations yield a minimum cost of 1.5. By deleting the first a in x, we shift the first sequence left, which aligns the subsequence a-c-b within both sequences. Then, by inserting b, we match the second b in y. Finally, substituting b for c, which costs less than inserting b and deleting c, we match the last b in y. The obtained distance is thus linked to the partially matched subsequence a-c-b-(bc), where (bc)—the substitution—reflects a partial match between states b and c. From a sociological point of view, the partially matched subsequence a-c-b-(bc) can be interpreted as a 'common backbone', or 'common narrative' between trajectories (Elzinga and Studer, 2013). In other words, the OM distance is based on the subsequence the two trajectories have in common.

Table 6. OM distance: Example

Sequence x	a	a	c	b	-	c	
Operations	d				i	s	
Sequence y	-	a	c	b	b	b	
Cost	0.5	0	0	0	0.5	0.5	Dist = 1.5

To summarise, an OM distance 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. Indel costs essentially allow us to control for admissible time warps in a sequence comparison, while substitution costs reflecting state dissimilarities serve as weights for the remaining mismatches. This is the general principle. The crucial question then becomes the choice of the indel and substitution costs.

3.3.2 Strategies for setting costs

We first examine methods for choosing substitution costs, and then discuss how to choose indel costs.

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

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. For example, full-time working looks closer to part-time working than to joblessness. The idea is then to set the costs to reflect the partial order stemming from this sort of theoretical background. This approach

Table 7. Two attributes of working statuses

Status	Responsibility Level	Jobless
Senior Manager (S)	4	no
Manager (M)	2	no
Employee (E)	1	no
Unemployed (U)	0	yes

Table 8. Substitution costs derived from state attributes using the Gower distance

	S	M	E	U
Senior Manager (S)	-			
Manager (M)	0.250	-		
Employee (E)	0.375	0.125	-	
Unemployed (U)	1.000	0.750	0.625	-

is used, for example, by Stovel et al. (1996) and McVicar and Anyadike-Danes (2002). To illustrate, assume careers coded using the following four statuses: Senior manager (S); Manager (M); Employee (E); and Unemployed (U). From the nature of the states, S is closer to M than to either E or U. 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 chosen for the ratios between the substitution costs remain quite arbitrary.

Costs based on state attributes A solution 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. 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 can use the Gower (dis)similarity coefficient (Gower, 1971).

Table 7 depicts the responsibility level and a joblessness indicator variable for the four working statuses. The Gower dissimilarities between the rows of the table are shown in Table 8.[6] These values can be used as substitution costs in an OM situation. Besides explicitly rendering the state comparison criteria, the approach also has the advantage of generating costs that surely satisfy the triangle inequality. Any concave transformation—such as the square root—of the distance between the states would also fulfil the triangle inequality.

[6]For the Gower distance between M and U, the contribution of responsibility level is the ratio between the difference and the range ($\frac{2-0}{4-0} = 0.5$), and that of jobless is 1 because the state is different. The resulting distance is $\frac{0.5+1}{2} = 0.75$.

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 for substituting between states when the transitions between them are rare, and a low cost when frequent transitions are observed. The transition rate between two states, a and b, is the probability $p(b \mid a)$ of switching from state a to state b between two successive positions. Assuming time invariance, the transition rate is estimated as

$$p(b \mid a) = \frac{\sum_{t=1}^{L-1} n_{t,t+1}(a,b)}{\sum_{t=1}^{L-1} n_t(a)} \quad , \tag{5}$$

where L is the maximum observed sequence length, $n_t(a)$ is the number of sequences with state a at position t that do not end in t, and $n_{t,t+1}(a,b)$ is the number of sequences with state a at position t and state b at position $t+1$. The symmetrical substitution cost is defined as[7]

$$\gamma_{tr}(a,b) = 2 - p(a \mid b) - p(b \mid a). \tag{6}$$

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. Practically, observed transition rates are generally low and the resulting substitution costs are all close to 2. Therefore, the OM distances based on transition-rate costs produce results close to those obtained 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. Such a generalisation is easily achieved by replacing $t+1$ with $t+q$, and $L-1$ with $L-q$ in the above equation. Whatever the time lag, q, the transition-rate-based substitution does not ensure the triangle inequality.

A conceptually better approach, in the spirit of the work by Rousset et al. (2012) described earlier, considers the two states a and b to be close when the chance that both states will be followed by a common state, c, q units of time later is high, i.e., when both states 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 high probability of leading to a managerial position, and a relatively low probability of leading to joblessness. We propose operationalising this idea by defining the substitution cost between a and b as the Chi-squared distance between the cross-sectional state distributions expected k time units after the occurrence of state a and state b,

$$\gamma(a,b) = \sum_{e \in \Sigma} \frac{\Big(p(e_{+q} \mid a) - p(e_{+q} \mid b)\Big)^2}{\sum_{f \in \Sigma} p(e_{+q} \mid f)}, \tag{7}$$

where $p(e_{+q} \mid f)$ is the probability of moving from f to e over q units of time.[8]

[7] Since we can arbitrarily substitute a to b in the first sequence or b to a in the second sequence, it would make sense to derive the cost from the highest transition rate only. In other words, we could also define $\gamma_{tr}(a,b)$ as $2(1 - \max\{p(a|b), p(b|a)\})$. This would give us values less close to 2.

[8] Using a negative k value, we can similarly determine costs in terms of a common past.

Another alternative to deriving costs from data was proposed by Gauthier et al. (2009). This approach is an 'optimisation' procedure based on methods used in biology (e.g. Henikoff and Henikoff, 1992). The principle is to consider two states as close—and assign them a low substitution cost—when they tend to jointly occur in pairs of similar sequences, and to consider them as dissimilar—and assign 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.

Indel costs. Despite the importance of indel costs for controlling time warp, choosing indel costs has, with the noticeable exception of Hollister (2009), received far less attention than substitution costs.

Single indel cost Indel is often seen as a gap insertion operator and so 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. Earlier, we established that, to obtain consistent distances, the extended $\mathbf{\Gamma}$ matrix of indel and substitution costs must fulfil the triangle inequality. This implies that a unique indel cost should not be smaller than half the maximum substitution cost. On the other hand, if we want indels to serve only to adjust sequence lengths, the fixed indel cost should be set equal to or greater than half the maximum substitution cost multiplied by the sequence length. Choosing an indel cost above the latter threshold would have no effect other than to severely penalise the differences in sequence lengths. For a fixed indel cost of c_I within the range:

$$\frac{\gamma_{\max}}{2} \leq c_I \leq L\frac{\gamma_{\max}}{2}, \tag{8}$$

where $\gamma_{\max}$ is the maximum substitution cost and L is the maximum sequence length, the combination of optimal indel and substitution operations will depend on the compared sequence patterns.

However, in some situations, we may be tempted to reduce the indel penalty below the lower threshold to allow for more time warp in a sequence comparison. For instance, Abbott and Tsay (2000) advocate using a low indel cost and suggest 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.

State-dependent indel costs Little attention has been paid to state-dependent indel costs, which is surprising because it would seem that the elements we delete and insert should be as important as elements we substitute. One of the rare applications that has used state-dependent indel costs is the already-cited study of sequences of yearly numbers of lynchings by Stovel (2001), where the indel cost is 1 for zero or one lynching,

and is otherwise equal to the number of lynchings. In this case, a greater number of lynchings is considered more of an exception, and the more exceptional a state, the more expensive it should be to insert or delete. This principle could serve as a general strategy for setting state-dependent indel costs. Like the resemblance between states, we can determine how exceptional a state is theoretically, based on the state's attributes, or derived from the data. As a data-driven solution, we propose defining 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 the 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 involved terms.

3.4 Variants of OM

Despite the high flexibility of OM with state-dependent costs, several authors (Elzinga, 2003; Hollister, 2009; Halpin, 2010; Elzinga and Studer, 2013) have pointed out that OM distances are essentially driven by differences in durations. There are two main reasons for that. First, sequences are, in social sciences, typically made of a few long spells and, therefore, the longest common subsequences typically include these longest spells or long portions of them (Elzinga and Studer, 2013). Second, OM operations are independently applied on each symbol in the sequence, regardless of the context. OM equally weights the insertion of state a in sequence aa and in sequence bb. In the first case, the insertion only affects the time spent in the spell in state a, whereas in the second case, it changes the sequencing (Hollister, 2009; Halpin, 2010). Lesnard (2010) observes that OM does not consider the position—age or date—when transformation operations are applied.

The OM variants 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. We do not discuss a proposition with similar aims by Dijkstra and Taris (1995) because, as shown by van Driel and Oosterveld (2001), the proposed algorithm does not produce the expected results.

Dynamic Hamming distance (DHD). State similarities in time-use analyses—for example, 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 generalised Hamming, and proposed that substitution costs should depend on the position t in the sequence.[9] He operationalises the idea by deriving the substitution cost at t from the transition rates cross-sectionally observed between $t-1$ and t and between t and $t+1$. The DHD time-dependent cost, $\gamma_t(a,b)$, at position t is thus

$$\gamma_t(a,b) = 4 - p_t(b \mid a) - p_t(a \mid b) - p_{t+1}(b \mid a) - p_{t+1}(a \mid b), \tag{9}$$

[9]Allowing for indels in addition to position-dependent substitution costs would make the problem far more complex by making the minimal cost for transforming one sequence into another depend on the order in which the edit operations are applied.

where $p_t(b \mid a)$ is the probability of switching from a to b between $t-1$ and t and is estimated as $n_{t-1,t}(a,b)/n_{t-1}(a)$.

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

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

$$c_I(z|a,b) = e\gamma_{\max} + g\,\frac{\gamma(a,z)+\gamma(b,z)}{2}, \tag{10}$$

where $\gamma()$ indicates the substitution cost, $\gamma_{\max}$ is the maximum substitution cost, and e and g are user-defined costs. The first term in Eq. 10 is similar to a fixed indel cost and e can be interpreted as a spell expansion cost or time-warp penalisation. The second term penalises differences with surrounding states, parameter g serving to control the importance of the penalisation. 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.

As long as the parameters e and g fulfil the constraint $1-2e \leq g$, the method also prevents the OM from using a pair of indels instead of a substitution, and thus provides a way to allow for important time warps while preserving the effectiveness of substitution costs. In her experiments, Hollister (2009) got the best results with a small shift penalisation, e, and a g close to $1-2e$.

Although she does not specify it in her paper, in the code she provided to us, Hollister computes the indel costs for each state at each position once at the beginning, and keeps them fixed during the alignment process. In other words, the indel costs remain unchanged, even after surrounding states are changed. This may generate dissimilarities that violate the triangle inequality, as shown in Table 9. It would make more sense to

Table 9. Localised OM distance between three sequences using a fixed substitution cost of 1, $e = .1$ and $g = .8$

	aabb	abbb	bbbb
x aabb	0.0		
y abbb	0.2	0.0	
z bbbb	2.0	1.0	0.0

adapt the indel cost after each operation. Consider sequences x and z in Table 9. Once

we have substituted a for b at the first position in z (cost=1), inserting a second a in front of the substituted a should cost 0.1. Then, deleting one of the last bs (cost=.1), we would obtain x for a total cost of 1.2 (instead of 2). In that case, the triangle inequality would be respected. However, adapting the indel cost after each operation raises computational issues since the total cost would vary with the order in which the successive transformation operations are applied. For instance, if we start by replacing the second b with an a in z, we would not end with the minimum cost of 1.2. Finding the minimum cost would require that we check all possible paths of transforming one sequence into the other, which quickly becomes intractable when the sequence length increases.[10]

By construction, the localised OM should be less sensitive than the classical OM to differences in spell length, while being more sensitive to changes in sequencing. Remember, however, that localised OM can generate dissimilarities that do not satisfy the triangle inequality.

OM sensitive to spell length. The localised OM distinguishes between indels that start/end a spell in a state and those that just expand or contract a spell. Moreover, it does not affect substitution costs. The OM variant proposed by Halpin (2010) accounts more explicitly for the spell length and makes indel and substitution costs depend on the spell length. The method can be formalised by distinguishing a state a in a spell of length t from a state a in a spell of length $t' \neq t$. Letting a_t denote any state a in a spell of length t, its indel cost $c_I^H(a_t)$ is determined as the basic indel cost c_I, corrected by a decreasing factor of t, and the substitution cost of, say, states a_{t_1} and b_{t_2} as the basic substitution cost $\gamma(a, b)$ multiplied by a factor decreasing with both t_1 and t_2. Using $1/t^h$ with $0 \leq h \leq 1$ as the exponent time weight, the costs are

$$c_I^H(a_t) = \frac{c_I(a)}{t^h}, \tag{11}$$

$$\gamma^H(a_{t_1}, b_{t_2}) = \gamma(a, b)\frac{1}{\max\{t_1^h, t_2^h\}}. \tag{12}$$

Halpin (2010) also suggests using an arithmetic or geometric mean of the inverse time length instead of the inverse of $\max\{t_1^h, t_2^h\}$ in Eq. 12. Whatever the solution retained for the exponent time weight, the resulting substitution costs, $\gamma(a_{t_1}, b_{t_2})$, could violate the triangle inequality and, therefore, generate unpredictable results.[11]

Decreasing the indel cost with the spell length produces the expected effect of favouring indels in longer spells instead of, for instance, indels which would create or suppress spells. The decrease of substitution costs with the lengths of the implied spells has, however, 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. Although the considered time lengths are not fixed a priori for all edit operations, as they are for the localised OM, there remains the question of whether we should consider the lengths of the spells before or after each indel or substitution operation.

[10] One other solution to render Hollister's method consistent could be to make substitution costs depend on surrounding states as well.

[11] For example, for $\gamma(a, b) = 1$ and $h = 1$, the substitution cost $\gamma^H(a_1, b_1)$ is 1, while we get the same substitution at a cost of .5 by first substituting a_2 to a_1 (cost 0) and then substituting b_1 to a_2 (cost .5).

OM between sequences of spells. For overcoming the limitations of the two former context-sensitive dissimilarities, we propose, in the spirit of the localised OM and the OM sensitive to spell lengths, to measure the OM distance between sequences of spells. The general idea is to consider, for each different value of t, a spell in state a during t units of times as a distinct element, denoted a_t, of the alphabet. Doing so considerably increases the alphabet size and, as a consequence, the number of indel and substitution costs to be considered. However, the number of parameters can easily be limited by expressing the cost $c_I^S(a_t)$ of the indel of spell a_t, as well as the substitution cost $\gamma^S(a_{t_1}, b_{t_2})$ between spells, a_{t_1} and b_{t_2}, in terms of the basic indel and substitution costs ($c_I(a)$ and $\gamma(a, b)$) of the constituting elements a and b and a correction factor function of the spell length. For instance, letting $\delta \geq 0$ be a weight factor for the spell length, the costs can be defined as

$$c_I^S(a_t) = c_I(a) + \delta \cdot (t - 1) \tag{13}$$

$$\gamma^S(a_{t_1}, b_{t_2}) = \begin{cases} \delta \cdot |t_1 - t_2| & \text{if } a = b \\ \gamma(a, b) + \delta \cdot (t_1 + t_2 - 2) & \text{otherwise.} \end{cases} \tag{14}$$

The parameter δ can be seen as the cost of extending or compressing a sequence by one unit of time, and the substitution between two spells $\gamma^S(a_{t_1}, b_{t_2})$ as the cost of compressing each spell into a one-unit-long spell, plus the substitution between the two concerned states, a and b. For $\delta < c_I(a)$, inserting an a in an existing spell a_t costs less than creating a new spell in a. The method therefore favours the expansion (or compression) of existing spells. Unlike the method by Halpin, it does not, however, encourage breaking long spells. Moreover, as defined by Equations 13 and 14, the costs $c_I^S()$ and $\gamma^S()$ satisfy the triangle inequality as long as $c_I()$ and $\gamma()$ verify the inequality. Interestingly, for $\delta = 0$, the OM of spell sequences becomes the OM distance between the sequences of DSS.

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

OM between sequences of transitions. Another way of accounting for the context, as described by Biemann (2011), is to compute the OM distances between the sequences of transitions. The transitions in a state sequence are characterised by its successive two long subsequences obtained by joining each state with its previous state, for example, the transitions in *aabb* are *aa*-*ab*-*bb*. Possibly, we could also specify the start of a sequence by a transition from the start to the first state and, likewise, the end of the sequence by a transition to the end.

As noted by Biemann (2011), by considering transitions instead of the states, we considerably increase the size of the alphabet and, hence, the number of indel and substitution costs to be considered. To overcome this limitation, we propose, similarly to what we have done for sequences of spells, to express the indel $c_I^B(a \to b)$ and substitution $\gamma^B(a \to b, c \to d)$ costs of transitions in terms of the indel and substitution costs of states. Considering that a transition $a \to b$ is made of an origin state a and a type of transition (e.g. a transition to the same state or transition to another state), we express the cost of inserting (substituting) a transition as a linear combination of the cost of inserting (substituting) the origin state and the cost $c_T(a \to b)$ of the concerned transition

type. Formally, we define the indel and substitution costs as follows:

$$c_I^B(a \to b) = wc_I(a) + (1-w)c_T(a \to b) \tag{15}$$

$$\gamma^B(a \to b, c \to d) = w\gamma(a,c) + (1-w)\Big(c_T(a \to b) + c_T(c \to d)\Big), \tag{16}$$

with $c_I(a)$ the (possibly normalised) indel cost of the origin state, a, $c_T(a \to b)$ the transition type cost, $\gamma(a,c)$ the (possibly normalised) substitution cost between the origin states a and c, and $w \in [0,1]$ a coefficient for controlling the trade-off between the cost related to the origin state and the cost related to the type of transition. A simple parameter-free solution for the $c_T(a \to b)$ function is to set it to 0 when $a = b$, and 1 otherwise. An alternative, which would make $c_T(a \to b)$ state dependent without the need for any additional parameters, is to set $c_T(a \to b)$ as the substitution cost, $\gamma(a,b)$, between a and b. Both of these solutions generate $c_I^B()$ and $\gamma^B()$ costs satisfying the triangle inequality when the basic costs $c_I()$ and $\gamma()$ 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 $w = 1$, and the measure shares the properties of OM in that case. Otherwise, by reducing w, we can increase the sensitivity to sequencing. Time warp can be controlled through the origin state indel cost, $c_I(a)$.

4 Recapping the dissimilarity measures

Given the multiplicity and diversity of dissimilarity measures surveyed above, it is worth summarising the different propositions and their relationships.

Table 10 recaps the characteristics of the main ways of measuring the dissimilarity between state sequences. The first column gives short names, which will be used later when presenting the results of our empirical evaluations. The next three columns correspond to the typology we adopted for the above overview, with 'Dis' denoting measures of differences between probability distributions, 'Att' measures based on the counts of common attributes, and 'Edt' the edit distances. The column 'Description' provides a short description.

The next five columns indicate the properties of the measures: 'Metric' denotes measures fulfilling the mathematical conditions of distances (required for most applications and especially for sample-based studies, as shown in Section 2), 'Eucl' for Euclidean distances (interesting property for multi-dimensional scaling), 'T.warp' for measures allowing for time warp in sequence comparison, 'S.dep' for state-dependent measures (i.e. for measures that allow for differences between states varying with the involved states), and 'Ctxt' for measures that consider the context of the states.

These properties may help to sharpen the set of potentially useful distances. For example, we may want to discard OM with the so-called optimised costs (OMopt) because of the possible negative values it can generate, but also non-metric measures such as OM with transitions-based costs (OMtrate), localised OM (OMloc), and dynamic Hamming (DHD), with costs derived from transition rates because of the unexpected behaviour that may result from the possible violation of the triangle inequality. Also, as already stated, non-Euclidean distances can lead to complex—non-real—coordinates in multidimensional

Table 10. Summary of dissimilarity measures between state sequences

	Type			Description	Properties					Parameters		
Measure	Dis	Att	Edt		Metric	Eucl	T.warp	S.dep	Ctxt	Subst.	Indels	Others
CHI2, EUCLID	x			Distance between per period state distributions	x	x	x					Number of periods K
CHI2fut (Rousset)	x			Position-wise state distances based on shared future	x	x			x			Time-lag weighting function
NMS (Elzinga)		x		Based on number of matching subsequences	x	x	x		x			
SVRspell		x		Based on number of matching spell subsequences with spell-length weights	x	x	x	x	x	User		Subsequence length weight a, spell duration weight b
HAM (Hamming)		x	x	Number of mismatches	x	x[b]						
generalized			x	Sum of mismatches with state-dependent weights	x[a]	x[b,c]		x		User		
DHD (Lesnard)			x	Sum of mismatches with position-wise state-dependent weights				x	x	Data		
OM			x	Minimum cost for turning x into y using theoretically defined costs	x[a]		x	x		User	Mult	
LCS / OM(1,2) / Levenshtein-II		x	x	Based on length of LCS / Number of indels	x		x					
feature			x	Costs based on state features	x		x	x		Features	Single	State features
future			x	Costs based on similarity between conditional state distributions q periods ahead	x		x	x		Data	Single	Forward lag q
trate			x	Costs based on transition rates			x	x		Data	Single	Transition lag q
opt[na](Gauthier)			x	Costs adjusted to increase similarity between similar sequences	[n]		x	x		Data	Single	Similarity rate
indels, indelslog			x	State dependent indels based on inverse or log inverse state frequencies.	x		x				Auto	
OMloc (Holister)			x	Context dependent indel costs			x	x	x	User	Auto	Expansion cost e, Context g
OMslen (Halpin)			x	Costs weighted by spell length	x		x	x	x	User	Mult[na]	Spell length weight h
OMspell (new)			x	OM between sequences of spells	x[a]		x	x	x	User	Mult[na]	Expansion cost e
OMstran (new)			x	OM between sequences of transitions	x[a]		x	x	x	User	Mult	Origin-transition trade-off w, Transition indel cost function

[a] If costs fulfil the triangle inequality. [b] Squared Euclidean distance. [c] If costs are squared Euclidean distances. [na] Not available in TraMineR. [n] Can generate negative dissimilarities.

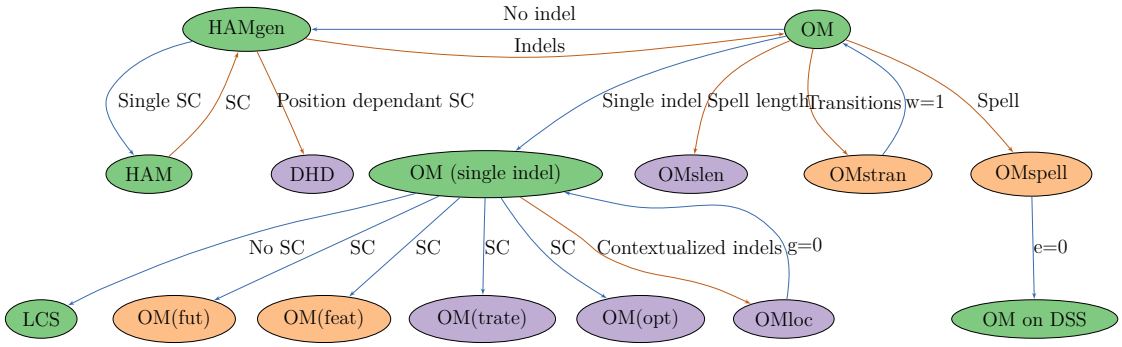


Figure 1. Relationships between OM variants. Node colours refer to the history of the measures: First-generation measures are green, second-generation measures are violet, and propositions made in this paper are orange. Red arrows represent extensions (with new aspects shown as labels) and blue arrows indicate specialisations (with corresponding constraint or parameter value given as labels). SC stands for substitution costs and SC alone refers to the specific way of setting values of state-dependent substitution costs.

scaling, and Euclidean distances may be preferred when finding principal coordinates is a concern.

The last columns in Table 10 show the available tuning parameters. 'Subst' stands for 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 ('Mult'), or if 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 the education level or the number of experienced childbirths, or result when some states share a higher number of common attributes than others. Possibility to consider state-dependent substitution costs is also of interest for the multichannel case. For example, the method adopted by Pollock (2007) for measuring distances between multichannel sequences consists of deriving the multichannel costs from the costs available for each individual channel, and generates costs that would at least vary with the non-matching channels.

A large majority of the surveyed dissimilarity measures are related to OM and it is instructive to look at how they are connected to each other. Figure 1 depicts the relationships between the many variants of OM distances. The figure reveals which dissimilarity measures generalise others, such as 'OMstran' (OM of sequences of transitions), which includes the classical OM as a special case, 'OMstran' being equivalent to 'OM' when the trade-off parameter w is set equal to 1, which puts all the focus on the transition origin state.

5 Simulation study

We have so far overviewed a great number of possibilities for measuring the dissimilarity between sequences. Table 10 lists no less than 11 types of distance measures and, for the OM distance, seven ways of setting the values of the costs. Moreover, many dissimilarity measures depend on user-defined parameter values and thus define families of measures. Varying parameter values provides some control on the behaviour of the measure. For instance, playing with the indel value, we can control the indel–substitution cost ratio of OM and, thereby, its sensitivity to timing. Altogether, we dispose of a huge number of possibilities of measuring dissimilarity and face the crucial question of choosing among them.

To help in that choice, this section provides empirical insights on how dissimilarity measures behave with regard to the three relevant aspects for comparing state sequences describing life trajectories, namely, sequencing, duration, and timing. We ran a series of simulation strands, and report the main outcomes below.

The reported simulations provide an original view of the ability of the dissimilarity measures to render differences in each of sequencing, duration, and timing dimensions, and differ in this way from other attempts to empirically compare dissimilarity measures. Several authors (see Robette and Bry, 2012, for a review) have analysed how results—most often the clusters derived from the dissimilarity values—change with the used dissimilarity measure. Such approaches permit us to assess the robustness of the outcome of the dissimilarity-based analyses against the used dissimilarity measure. However, outcome-oriented simulation analyses do not stricto-sensu provide indications on the behaviour of the measures, and the generalisation of their findings to other data sets and analysis methods—clustering algorithms—is subject to debate. The approach by Robette and Bry (2012), based on correlations between dissimilarities computed on artificial data is more illuminating from that point of view. Nevertheless, while the Mantel tests of correlations used by those authors prove useful in identifying measures that behave similarly, they do not say to what the measures are sensitive.

5.1 Simulation design

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

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 evaluated characteristic—sequencing, duration, or timing—is kept fixed for all sequences in the group, while the other aspects are randomly changed across the sequences so as to also allow for non-systematic differences between sequences on those other non-evaluated aspects. Doing so, the compared sequences differ systematically in the evaluated aspect, but also randomly differ on all other aspects. We can thus evaluate the relative importance given by the dissimilarity measures to the selected aspect in the presence of discrepancies on the others, and, hence, evaluate how good each of the measures is in rendering differences of the studied aspect.

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

The sensitivity to the considered criterion is measured with the pseudo R^2 defined in Studer et al. (2011) for measuring 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 a strong sensitivity to the considered systematic difference between the two groups. In other words, it is able to discriminate between the groups. In contrast, a low R^2 value indicates that the measure poorly accounts for the tested dimension. In order to ensure stable R^2 values, $1,000,000$ of sequences per group were generated in each series of simulations. All simulated sequences are of length 20.[12]

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. They are not comparable, however, between series or strands, since the total variability and the mean R^2 differ significantly across series. We therefore report the standardised value, namely, the score S_d defined as the standardised R^2_d given by Equation (17), where $\overline{R^2}$ and s_{R^2} are, respectively, the mean value and standard deviation of the $R^2_{d_i}$s in the series.

$$S_d = \frac{R^2_d - \overline{R^2}}{s_{R^2}}. \tag{17}$$

The score reflects the sensitivity of each dissimilarity measure, d, in comparison with the overall sensitivity of all considered measures. The score is positive for dissimilarity measures more sensitive than the average to the tested dimension, and negative otherwise.

5.2 Random sequence generation

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

For clarity and owing to space constraints, we report only a subset of all the—sometimes more sophisticated—series of simulations we tried (see Studer, 2012). However,

[12]Despite the huge number of sequences, all computations could be done relatively quickly by considering only unique sequences and weighting them by their counts. Each set of 1,000,000 simulated sequences contained between 80 and 800 unique sequences, except for one set that had around 3,000 unique sequences.

the experiments reported can be considered representative in that they render all salient findings of the complete set of tried simulations.

5.2.1 State-based generating process

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

We report three strands of state-based simulations whose characteristics are summarised in Table 11. 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. The second strand evaluates the sensitivity to timing. The order patterns are randomly selected and durations randomly set, while controlling the start of the spell in state c in each of the two groups. Several series of simulations have been run by varying the difference in the start time in c between the two groups. Finally, the third strand evaluates the sensitivity to duration by controlling the total consecutive time spent in a given state for each group.

Table 11. Designs for evaluating sensitivity to order, timing, and duration of states

Tested Dimension	Description	Group One	Group Two
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 start time t of the spell in state 'c' is controlled.	$t = 7$ $t = 15$	$t \in \{9 \dots 15\}$ $t \in \{7 \dots 13\}$
Duration	Sequences randomly follow one of the patterns 'abc' or 'cba' and duration d of the spell in state 'b' is controlled.	$d = 4$ $d = 14$	$d \in \{6 \dots 14\}$ $d \in \{4 \dots 12\}$

We ran an additional strand of simulations to evaluate the sensitivity of the measures to small perturbations (see Table 12). 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.

Table 12. Designs for evaluating sensitivity to a random change of state

Description	Order Pattern	State Change in Group Two
Controlled order pattern and random durations. Sequences in second group derived from the sequences of the first group by randomly changing one of their elements.	'abc'	anywhere
	'abc' or 'cba'	anywhere
	'abc'	start/end of spells
	'abc' or 'cba'	start/end of spells

5.2.2 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. For example, let l be leaving home, m be marriage, and c be first childbirth. Before any of those three events occurs, we are in the initial state i (living with parents). Then, depending on the first event experienced, we switch to one of the states 'having left home', 'being married', or 'having a child', which we denote using the symbols of the corresponding events l, m, or c. State m therefore means 'married without having left home'. When the second event occurs, we would, in the same way, switch to one of the states lm, lc, or mc, where lm, for instance, means 'having left home and married'. Eventually, after all three events have occurred, we would be in state lmc.

For our simulations, we consider three events. Each sequence is then characterized by the occurrence times of the three events e_1, e_2, e_3. The sequences are simulated by generating the occurrence times 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 one aims at evaluating the sensitivity to the order of occurrences of the events. We impose that the first event occurs before the second one in group 1, and after the second one in group 2. The second strand evaluates the sensitivity to the timing of the events by controlling the occurrence time of event 1 in each group. To evaluate sensitivity to the duration between two events, we control, in the last strand, the elapsed time between the first two events. The occurrence time of the third event is left completely random in all cases. Table 13 summarizes the three strands of simulations based on event occurrences.

When comparing event-based sequences, we can, for state-dependent measures, define the state dissimilarities—substitution costs—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.

5.3 Analysed dissimilarity measures

Most of the dissimilarity measures recapitulated in Table 10 have been included in the simulation study. For distances that can be parameterized, we have considered a selection

Table 13. Simulations evaluating the sensitivity to the order and timing of events, and to duration between events

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

of parameter configurations so as to explore the effect of the parameters and the range of behaviours that can be covered. The complete list of dissimilarity measures and parameter configurations studied in the simulations is given in Table 15, and the meaning of the parameters used in that table and in the following figures is specified in Table 14.

Table 14. Meaning of parameters used in Table 15

Label	Description
K	Number of intervals used to compute Chi-square and Euclidean distances.
i	Indel cost (single cost of 1 when not specified)
sm*	Substitution costs: (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 vs. transition-type trade-off
ti	Transition indel costs: (single cost of 1 when not specified), sm (based on substitution costs), raw (Bienmann'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 $h = .5$)

* To save space, 'sm=' will be omitted and therefore OM arguments without the '=' sign should be interpreted as values of the sm argument.

Table 15. 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)
Optimal Matching (OM)	OM, OM(i=1.5), OM(trate), OM(indelslog), OM(indels), OM(future)
Localized Optimal Matching (OMloc)	OMloc(e=0, 0.1, 0.25, 0.4)
Spell-Length-Sensitive Optimal Matching (OMslen)	OMslen(h=1, i=1, 1.5, 5), OMslen(i=1, 1.5, 5)
Optimal Matching of Spell Sequences (OMspell)	OMspell(e=0, 0.1, 0.5, 1), OMspell(e=0, 0.1, 0.5, 1, i=2)
Optimal Matching of Transition Sequences (OMstran)	OMstran(w=0, 0.1, 0.5), OMstran(i=1.5, w=0.1, 0.5), OMstran(i=5, 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)

5.4 Results

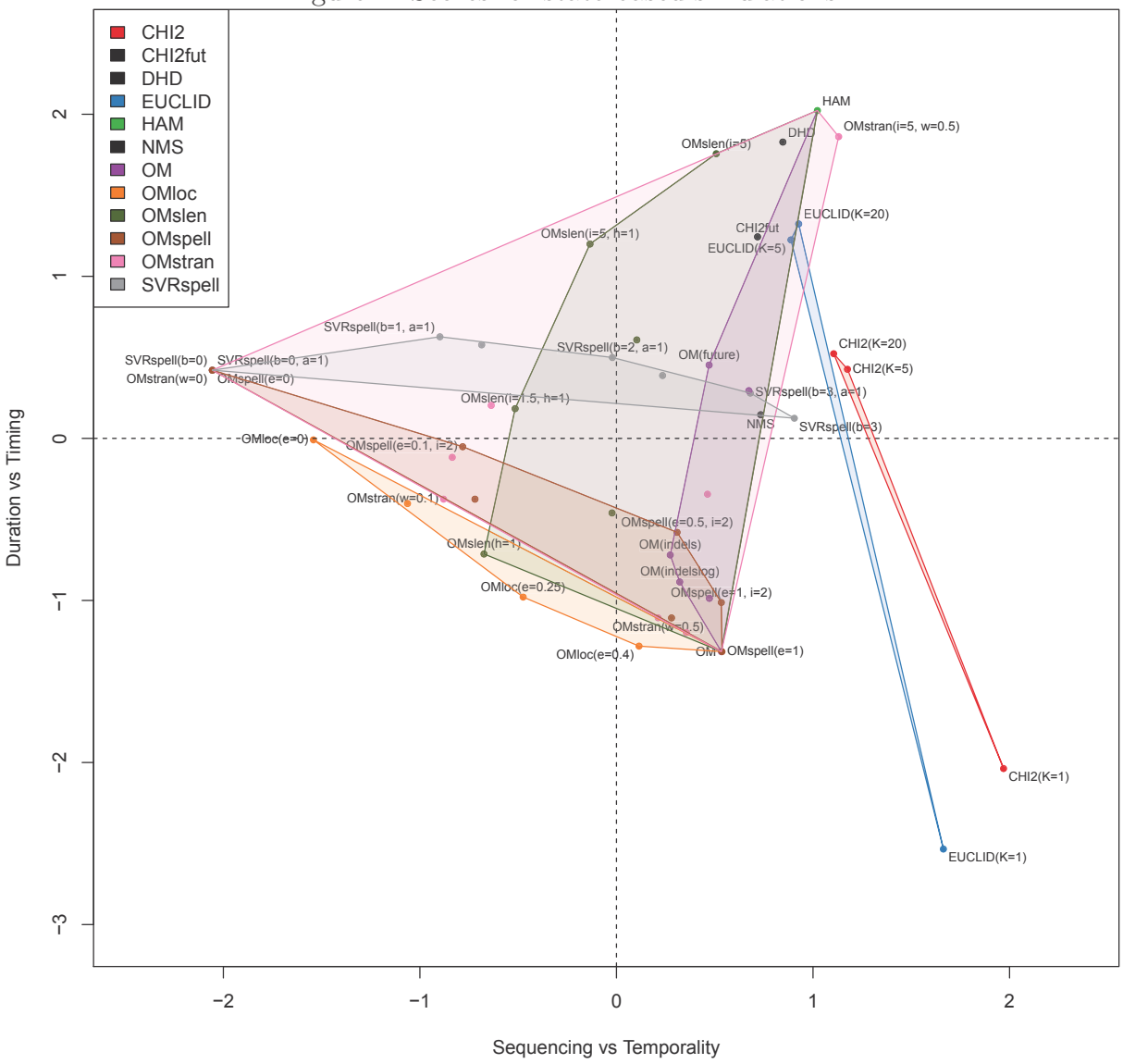
Detailed results for each series of simulations are provided as an online appendix. We summarize here the outcome of the study by opposing the mean scores achieved by each measure for the simulations of the 'sequencing' strand to the mean scores obtained for the temporality—'timing' and 'duration'—strands[13] on one side and the 'duration' scores to the 'timing' scores on the other. These two axes roughly correspond to the first two robust principal components found in Studer (2012). Unlike principal components, however, the axes here are defined independently from the data. They also have a clearer interpretation. The first axis is defined as temporality score minus mean sequencing score and will, therefore, be oriented such that higher sensitivity to sequencing is on the left and higher sensitivity to temporality dimensions on the right. The second axis is defined with higher sensitivity to durations at the bottom and higher sensitivity to timing at the top.

Results from the state-based group of simulations are graphically displayed in Figure 2 and the results for the event-based group are in Figure 3. Figure 4 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 other ones and does not reflect any absolute level of sensitivity.

In order to not overload the figure with too many points, the results for each family of measures is represented by the smallest polytope covering the scores obtained for the different tested parameterisations. The labels of inner points are omitted and only those of configurations associated with the vertices of the polytope are displayed. A large

[13]The mean temporality scores are computed as the average between the mean timing and mean duration scores. The reported mean scores have been standardized.

Figure 2. Scores for state-based simulations

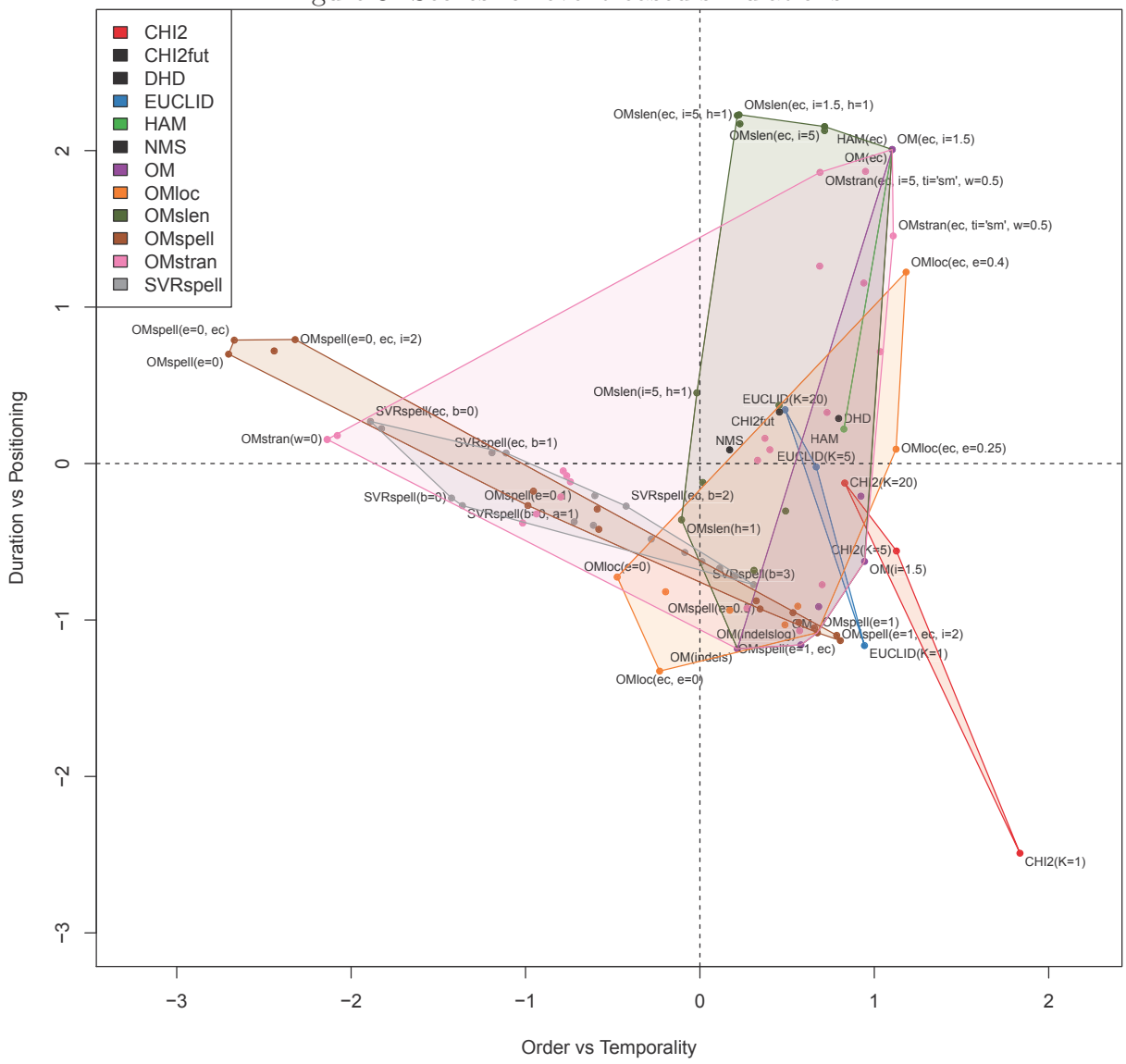


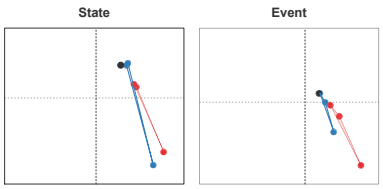
polytope area, such as the one for OMstran—OM of transitions—in Figure 2 indicates that the measure allows for very different sensitivities through its parameterisation.

We can observe that the measures are distributed within a triangle in Figures 2 and 3. This (unsurprisingly) reflects a higher contrast between duration and timing sensitivities among measures sensitive to temporal aspects—on the right—than among measures primarily sensitive to the sequencing—on the left. A noticeable general outcome in Figure 3 is that taking account of the explicit information on state proximities can significantly affect the behaviour of the measure, e.g., HAM(ec) lies very far from HAM.

Results by distance families Let us have a closer look at each considered family of dissimilarity measures. Two small graphs give the position each family occupies in Figures 2 and 3.

Figure 3. Scores for event-based simulations

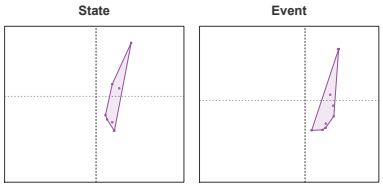




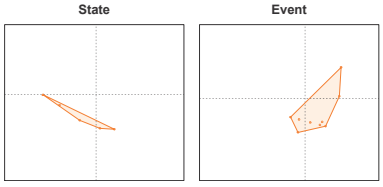
Distribution-based distances: Unsurprisingly, dissimilarity measures based on differences between distributions over the whole period (K=1) appear to be the most sensitive to durations. They also are the least sensitive to differences in sequencing with R^2s close to 0. The Chi-square distance CHI2(K=1) is, among all considered distances, the most sensitive to duration differences for rare states, while the Euclidean distance EUCLID(K=1) is the most sensitive to differences for states with high durations. When K increases, the sensitivity of the CHI2 measure shifts from duration to timing. For K equal to the sequence length (here, 20), CHI2 gets scores similar to those of the Hamming family regarding timing, but maintains some sensitivity to differences in durations. The detailed results in the online appendix exhibit that the position-wise Chi-square distance ranks better as a time-sensitive measure for small time changes than

for large differences in timing. CHI2fut—itself a position-wise measure—is closer to the position-wise CHI2 as compared to CHI2 versions with a smaller K.

Hamming: All variants of the Hamming distance lie in the top-right quadrant, meaning they are specifically sensitive to timing differences. They are slightly less insensitive to differences in sequencing than overall distribution-based distances. This is because sequencing is partly determined from the starting and ending 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 reached by HAM(ec) using event-based substitution costs. The HAM and DHD scores are low relatively to the score of HAM(ec). From the simulations, the time varying costs of the DHD metric seem to relax somewhat the strong time sensitivity of pure Hamming. As with the position-wise Chi-square distances, the Hamming distances rank better as time-sensitive measures for small time changes than for large time differences.

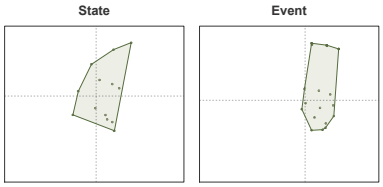


Optimal matching: The family of OM distances lies on the right of the plot, which confirms the low sensitivity to differences in sequencing, as pointed out, for instance, by Elzinga (2003), Hollister (2009), and Halpin (2010). As expected, we also observe that OM with high indel—relatively to substitution 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. The scores for OM with data driven substitution costs remain, as expected, very close to the scores from 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) 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 in the top-right quadrant. As also expected, deriving substitution and indel costs from the state frequencies renders OM more sensitive to changes in the duration of rare events (Figure 4). Using the log of inverse frequencies seems a better choice than raw inverse frequencies that make OM too sensitive to rare events and small perturbations. Costs derived from the count of non-shared lived events (ec) reduce the sensitivity to duration and ensure more importance to timing differences. Interestingly, in all our simulations, OM(ec) and OM(ec, i=1.5) get the same scores as HAM(ec), the reason being that the substitution costs are globally so low that indels costing 1 or more are never used.

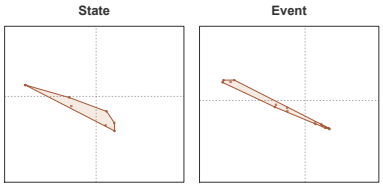


Localized optimal matching: The distances of the localized OM family are, with a few exceptions in the case of costs based on the count of non-shared events, situated in the lower portion of the plots. The horizontal position is determined by the expansion cost e: the lower the value of e, the more to the left is the position, with the measure becoming highly insensitive to temporality when e approaches 0. For some simulation

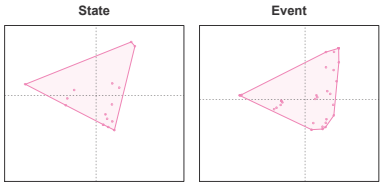
strands, the part of R^2 for timing differences that is accounted for by the measure, becomes negative. This means that with OMloc, we can get a total within-group discrepancy greater than the overall 'OMloc' discrepancy. This is a consequence of the violation of triangle inequality and makes OMloc especially unsuited for distinguishing between groups of sequences with different timings.



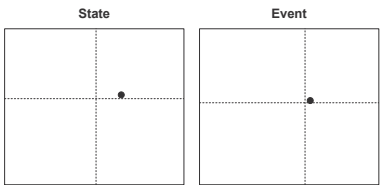
Spell-Length-Sensitive Optimal Matching: As expected by Halpin (2010), OMslen appears to be less sensitive to differences in durations than classical OM. However, here again, for several simulation strands (related to timing, duration, and random perturbation) we got negative R^2s when $h = 1$.



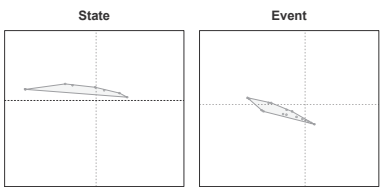
Optimal Matching of Spells: The family of OMspell distances is located around a line going right to left from OM—classical OM with a single substitution cost—to OM of the DSS sequences, the latter corresponding to OMspell (e=0). A high expansion cost e makes the measure more sensitive to temporality, while low values of e give more importance to sequencing. The sensitivity to temporality is attributable primarily to duration rather than timing.



Optimal Matching of Transitions: The family of OM distances between sequences of transitions is the one that covers the largest range of sensitivity combinations. Sensitivity to temporality increases with the value of the origin-transition trade-off parameter w. Recall that for $w = 1$, OMstran is equivalent to classical OM. Lowering the value of w significantly increases sensitivity to sequencing. As with classical OM, the vertical position is mainly driven by the indel/substitution cost ratio. We can observe, however, that the effect of the ratio becomes smaller when w decreases, i.e., when more importance is given to the transition type than to the origin states.



NMS: The NMS distance occupies a neutral position near the centre of the plots. While such neutral positions result in other distances exhibiting balanced positive sensitivity to sequencing, temporality, and duration, this is not the case for NMS, which appears to be more or less unsensitive to each of them. In Figure 4, for instance, we can observe that NMS is the least sensitive distance to ordering. Counter-intuitively, the measure gets 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 on the sequences of DSS, which corresponds to SVRspell ($b = 0$).



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. 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 contrarily 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. This behaviour is not confirmed, however, by the event-based simulations. The effect of the a parameter that determines the weight given to the length of the subsequences remains unclear but limited. Looking at Figure 4, we observe that the SVRspell measures are, after the CHI2 measures, the most sensitive to small random perturbations.

Small random perturbations The sensitivity to small random perturbations is shown in Figure 4 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 Chi-square distances are, regardless of the breakdown of the covered time interval into periods, quite insensitive to differences in ordering. The Chi-square distance appears to be, in our simulations, much more sensitive to small perturbations than the basic Euclidean distance. If we exclude the Chi-square distance, we observe that parameterisations that make measures more sensitive to ordering, render them—at the same time—more sensitive to small perturbations. The linear correlation between the ordering scores and the scores for random perturbation of all but Chi-square measures is .8.

6 Choosing the right dissimilarity measure

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

The previous discussion allows us, however, to discard a few measures. These are the NMS distance of Elzinga (2003), OMloc (the localized OM) of Hollister (2009), and OMslen (the spell-length-sensitive OM) of Halpin (2010). NMS lacks sensitivity to all three aspects, OMloc has strange behaviour resulting from violation of the triangle inequality, and OMslen has counter-intuitive and hence unexpected behaviour. Although these three measures have interesting characteristics, alternatives—SVR for NMS, OMstran for OMloc, and OMspell for OMslen—sharing similar aims without suffering from their drawbacks should be preferred. Further, remember that we discarded the so-called 'optimisation cost' method 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 very close from OM with a single state-

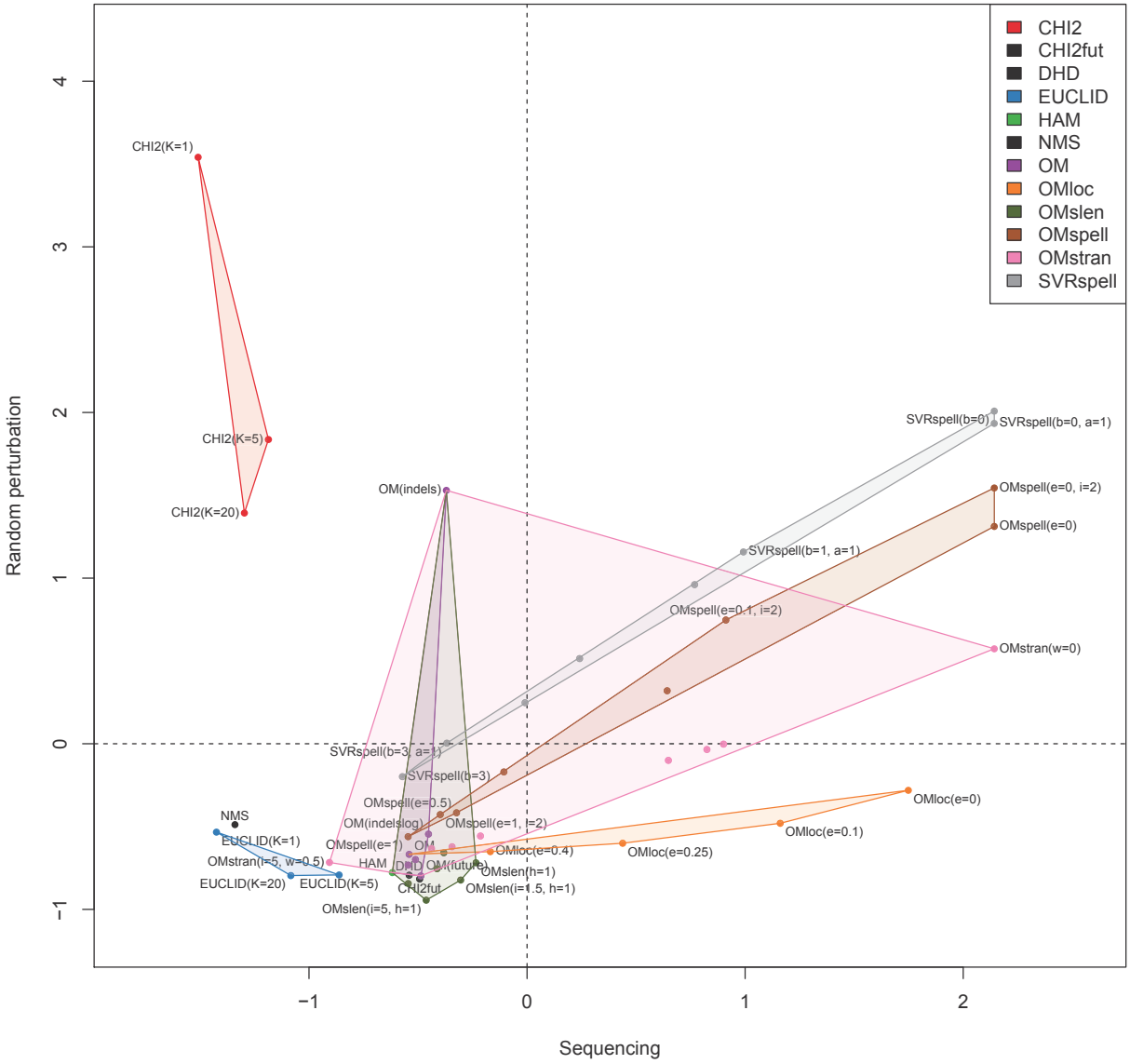


Figure 4. Sensitivity to a random change of state vs. sensitivity to sequencing

independent substitution cost.[14] The same holds for DHD, which produces results similar to simple Hamming (HAM). Together with the questionable justification for linking substitution costs to transition rates, the above remarks advocate against using such transition-rates-based methods.

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

If the focus is on changes in sequencing, measures highly sensitive to sequencing should

[14]In contrast, when states are structurally organised as in our event-based simulations, taking this information on the relationships between states into account—the ec cases—can drastically change the outcomes.

be preferred. Good choices are OMstran—OM of transitions—with low weight (low w value) on the state of origin, OMspell—OM of spells—with low expansion cost e, and SVRspell—subsequence vectorial representation metric of Elzinga and Studer (2013)—with low b spell length weight. One of the differences between these three measures is the sensitivity to small perturbation. If one is interested in these small differences, such as small spells of unemployment for instance, SVRspell should be used. On the contrary, OMstran appears to be less sensitive to this aspect, whereas OMspell shows an intermediary position. Classical OM is definitively not suited for measuring differences in sequencing.

If we are interested to explain changes in timing, then we need measures sensitive to timing. Position-wise measures such as those of the Hamming family are the most time sensitive. Using the CHI2 (Deville and Saporta, 1983) 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 indeed can also be achieved by expressing the Hamming distance as an OM distance with high indel and progressively lowering the high indel value. CHI2 is specifically interesting when we want to give high importance on changes involving rare states.

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

While 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 simultaneously take into account differences in two or three dimensions out of timing, duration, and sequencing. Measures such as OMstran, OMspell, and SVRspell—that can cover a large number of different mixes of sensitivities—look interesting in this multi-focus context for the control they allow on the trade-off between the different dimensions.

In many applications, we may be interested in the specific differences attributable to each of the timing, duration, and sequencing aspects rather than considering them simultaneously. It could then be useful to use three different dissimilarity measures: one sensitive to timing, one to duration, and one to sequencing. It should then be possible to identify distinctions stemming from each aspect by comparing the analysis outcomes obtained with each of the considered measures. For example, when studying the long-term consequences of professional insertion 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 impact 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 find out whether differences are primarily due to changes in sequencing, thereby reflecting, for instance, the emergence of new stages such as 'cohabiting couples', or due to timing differences resulting from the postponement of events such as marriage or childbirths, or due to differences in spell durations such as marriage duration or the

delay between marriage and first childbirth.

Running cluster analyses with different dissimilarity measures should also allow us to find out 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 obtained. In a discrepancy analysis, comparing outcomes obtained with different measures may also help to identify which covariates explain best sequencing differences, and which ones explain best timing and duration differences.

It is worth recalling, however, that the different dimensions are not completely independent from each other. We may therefore observe only minor differences between outcomes obtained with different measures. Nevertheless, the use of multiple measures can provide interesting knowledge about borderline cases, and we could learn, for instance, that a given trajectory looks more like a type A regarding sequencing and more as a type B from the timing point of view.

7 Conclusion

Sequence analysis, in particular, dissimilarity-based sequence analysis, has gained much popularity in life course studies in the past few years. Although OM remains the most used dissimilarity measure, there exists many other ways of measuring dissimilarity that have either been developed independently or have been specifically proposed to answer criticisms addressed to OM. The end user faces, therefore, the difficult task of choosing a suitable measure for her/his research objectives. The structured and critical review of the literature proposed in this article together with the simulation study of the behaviour of the many addressed variants are meant to help one make this choice. The review is original in several respects. Firstly, it is specifically oriented towards the ability of the measures to render timing, duration, and sequencing differences that matter for life course studies. Secondly, 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 have unexpected behaviours. Thirdly, the review covers a unique list of measures including novel alternatives to fix weaknesses of existing measures or to fill a lack of existing measures able to consistently take a specific aspect into account.

A few aspects not addressed in the paper due to space limitations 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 normalised distances. There exists evidence regarding these aspects. Measures which do not allow time shift in sequence comparison clearly are not applicable on sequences of different length. It is also known that normalisation can break mathematical properties such as the fulfilment of the triangle inequality (Gabadinho et al., 2011, p 29). 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 similar to the one in this article.

To conclude, it is worth mentioning that the entire set of dissimilarity measures studied with simulated sequences have been implemented in the TraMineR R package. Appendix A shortly describes the function that computes them and draws the links between

the arguments of the function and the distance and parameter notations used in this article.

Acknowledgments:

This publication results from research work executed within the framework of the Swiss National Centre of Competence in Research LIVES (IP14), which is financed by the Swiss National Science Foundation. The authors are grateful to the Swiss National Science Foundation for its financial support.

A Computing the dissimilarities with TraMineR

Most of the dissimilarity measures considered in this article have been made available in the latest version of the TraMineR package (Gabadinho et al., 2011), a library for sequence analysis in R. We shortly describe here how to compute distances between sequences with the `seqdist` function provided by this package.

To use `seqdist`, we first have to define a state sequence object with the `seqdef` function. We illustrate using the `mvad` data set that ships with TraMineR. For each case, the successive states forming the sequences are stored in columns 17 to 86 of the `mvad` data frame. We load the library and data, and create the state sequence object with

```
R> library("TraMineR")
R> data("mvad")
R> mvad.seq <- seqdef(mvad[,17:86])
```

The matrix `dis` of the pairwise dissimilarities between the sequences is then obtained with

```
R> dis <- seqdist(mvad.seq, method = distancename, other arguments)
```

where `distname` can be any of the measure names listed in Table 10.

The minimal arguments are the state sequence object and the method. Depending on the method selected, additional arguments may be required to specify values of parameters. There also are optional arguments to ask for normalisation, take weights into account, or specify a reference sequence from which to compute dissimilarities for example. Table 16 summarizes the available arguments.

Table 16. Arguments of TraMineR's `seqdist` function

Argument	Value	Description
		Generic
`with.missing`	Logical	If `TRUE`, missing values are coded as an additional state. Since some methods do not handle missing values, `TRUE` may be required when the sequences contain missing values.
`refseq`	see description	Optional baseline sequence to compute the distances from. Can be the index of a sequence in the state sequence object (0 for the most frequent sequence), or an external sequence passed as a sequence object with a single row.
`norm`	Character: `''none''`, `''maxlength''`, `''gmean''`, `''maxdist''`, `''YujianBo''`	Normalisation method (default is `''none''`).
`weighted`	Logical	Should weights be taken into account? (Ignored when not applicable.)
`full.matrix`	Logical	If `TRUE` (default), the full distance matrix is returned. This is for compatibility with earlier versions of the `seqdist` function. If `FALSE`, an object of class `dist`, i.e., a vector with only values from the upper triangle of the distance matrix, is returned. Since the distance matrix is symmetrical, no information is lost with this representation while size is divided by 2. Objects of class `dist` can be passed directly as arguments to most clustering functions. Ignored when `refseq` is set.
`sm`	Character or matrix	Substitution-cost matrix specifying distances between states. Can also be `''TRATE''` (transition-rate-based) or `''CONSTANT''` (single substitution cost).
		Chi-square and Euclidean (`method=''CHI2''` or `method=''EUCLID''`)
`breaks`	Numeric	An optional vector specifying the period limits.
`step`	Numeric	Length of fixed-length periods (ignored when `breaks` is not `NULL`).
`overlap`	Logical	if `TRUE`, overlapping intervals are built using the `step` value (ignored when `breaks` is not `NULL`).
		Hamming (`method=''HAM''`)
		Dynamic Hamming (`method=''DHD''`)
`sm`	3-dimensional array	Substitution-cost matrix specifying distances between states at each time point. The third dimension is used for the position in the sequences. If omitted, substitution costs are based on transition rates.
		Optimal Matching (`method=''OM''`)
`indel`	Vector or single value	Cost for inserting/deleting a state. Either one value or a vector with a value for each element of the alphabet.
		Localized Optimal Matching (`method=''OMloc''`)
`expcost`	Numeric	Cost of expending/contracting the spell length.
`context`	Numeric	Penalisation for differences with surrounding states.
		Duration Sensitive OM (`method=''OMslen''`)
`h`	Numeric	Exponent of spell length
`link`	Character	Function to compute substitution cost: `''mean''` (arithmetic average, default) or `''gmean''` (geometric mean), or `''max''` (maximum).
`indel`	Vector or single value	Basic indel costs.
		OM between sequences of spells (`method=''OMspell''`)
`indel`	Vector or single value	Basic indel costs.
`expcost`	Numeric	Cost of expending/contracting the spell length.
		OM between sequences of transitions (`method=''OMstran''`)
`indel`	Vector or single value	Basic state indel costs.
`transindel`	Character	Method to compute transition indel costs: either `''constant''` (all equal to 1) or `''subcost''` (based on substitution costs)
`otto`	Numeric (0-1)	Weight for controlling the trade-off between cost of origin state and cost of transition type.
`previous`	Logical	If `TRUE`, add transition from previous state.
`addcolumn`	Logical	If `TRUE`, replicate first (when `previous=TRUE`) or last column.
		Number of Matching Subsequences (`method=''NMS''`)
`sm`	Matrix	State *proximities* (not distances!).
		Sequence Vectorial Representation Metric (`method=''SVRspell''`)
`sm`	Matrix	State *proximities* (not distances!).
`a`	Vector or single value	Exponent weight of subsequence length
`b`	Numeric	Exponent weight of spell length

References

Abbott, A. (1983). Sequences of social events: Concepts and methods for the analysis of order in social processes. *Historical Methods*, 16(4):129–147.

Abbott, A. (1990). A primer on sequence methods. *Organization Science*, 1(4):375–392.

Abbott, A. (2000). Reply to Levine and Wu. *Sociological Methods & Research*, 29(1):65–76.

Abbott, A. and Forrest, J. (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History*, 16:471–494.

Abbott, A. and Hrycak, A. (1990). Measuring resemblance in sequence data: An optimal matching analaysis of musician's carrers. *American Journal of Sociolgy*, 96(1):144–185.

Abbott, A. and Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology, Review and prospect. *Sociological Methods and Research*, 29(1):3–33. (With discussion, pp 34-76).

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. *Sociological Methods and Research*, 38(3):430–462.

Bergroth, L., Hakonen, H., and Raita, T. (2000). A survey of longest common subsequence algorithms. In *String Processing and Information Retrieval, 2000. SPIRE 2000. Proceedings. Seventh International Symposium on*, pages 39–48.

Biemann, T. (2011). A transition-oriented approach to optimal matching. *Sociological Methodology*, 41(1):195–221.

Billari, F. C., Fürnkranz, J., and Prskawetz, A. (2006). Timing, sequencing, and quantum of life course events: A machine learning approach. *European Journal of Population*, 22(1):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. *Journal of Econometrics*, 22:169–189.

Dijkstra, W. and Taris, T. (1995). Measuring the agreement between sequences. *Sociological Methods and Research*, 24(2):214–231.

Elzinga, C. H. (2003). Sequence similarity: A non-aligning technique. *Sociological Methods and Research*, 31:214–231.

Elzinga, C. H. (2005). Combinatorial representations of token sequences. *Journal of Classification*, 22:87–118.

Elzinga, C. H. (2007). Sequence analysis: Metric representations of categorical time series. Manuscript, Dept of Social Science Research Methods, Vrije Universiteit, Amsterdam.

Elzinga, C. H. and Studer, M. (2013). Spell sequences, state proximities and distance metrics. *Sociological Methods and Research*. Revise and Resubmit.

Gabadinho, A. and Ritschard, G. (2013). Searching for typical life trajectories applied to childbirth histories. In Levy, R. and Widmer, E., editors, *Gendered life courses - Between individualization and standardization. A European approach applied to Switzerland*, pages 287–312. LIT, Vienna.

Gabadinho, A., Ritschard, G., Müller, N. S., and Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4):1–37.

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. *Sociological Methods and Research*, 38:197–231.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27:857–874.

Gower, J. C. (1982). Euclidean distance geometry. *Mathematical Scientist*, 7:1–14.

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. *Sociological Methods and Research*, 38(3):365–388.

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal*, 26(2):147–160.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–10919.

Hogan, D. P. (1978). The variable order of events in the life course. *American Sociological Review*, 43:573–586.

Hollister, M. (2009). Is Optimal Matching Suboptimal? *Sociological Methods Research*, 38(2):235–264.

Kruskal, J. B. (1983). An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM Review*, 25(2):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 socio-temporal patterns. *Sociological Methods and Research*, 38:389–419.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710.

Levine, J. (2000). But what have you done for us lately. *Sociological Methods & Research*, 29 (1):pp. 35–40.

Liefbroer, A. C. and Elzinga, C. H. (2012). Intergenerational transmission of behavioural patterns: How similar are parents' and children's demographic trajectories? *Advances in Life Course Research*, 17:1–10.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins, C. (2002). Text classification using string kernels. *The Journal of Machine Learning Research*, 2:419–444.

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 International Workshop, WSOM 2009, St. Augustine, FL, USA, June 8-10, 2009*, volume 5629 of *Lecture Notes in Computer Science*, pages 154–162. Springer, Berlin.

McVicar, D. and Anyadike-Danes, M. (2002). Predicting successful and unsuccessful transitions from school to work using sequence methods. *Journal of the Royal Statistical Society A*, 165(2):317–334.

Needleman, S. and Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 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, CNRS Marseille and PAVIE University of Lausanne. (available at http://www.tcoffee.org/saltt/).

Pollock, G. (2007). Holistic trajectories: A study of combined employment, housing and family careers by using multiple-sequence analysis. *Journal of the Royal Statistical Society A*, 170(1):167–183.

Robette, N. and Bry, X. (2012). Harpoon or bait? a comparison of various metrics in fishing for sequence patterns. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 116(1):5–24.

Rousset, P., Giret, J.-F., and Grelet, Y. (2011). Les parcours d'insertion des jeunes: une analyse longitudinale basée sur les cartes de Kohonen. Net.Doc 82, Céreq.

Rousset, P., Giret, J.-F., and Grelet, Y. (2012). Typologies de parcours et dynamique longitudinale. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 114(1):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, WOG Historical Demography, Leuven.

Settersten, Richard A., J. and Mayer, K. U. (1997). The measurement of age, age structuring, and the life course. *Annual Review of Sociology*, 23:233–261.

Stovel, K. (2001). Local sequential patterns: The structure of lynching in the deep south, 1882-1930. *Social Forces*, 79(3):pp. 843–880.

Stovel, K., Savage, M., and Bearman, P. (1996). Ascription into achievement: Models of career systems at lloyds bank, 1890-1970. *American Journal of Sociology*, 102(2):358–399.

Studer, M. (2012). *Étude des inégalités de genre en début de carrière académique à l'aide de méthodes innovatrices d'analyse de données séquentielles*, volume SES-777 of *Collection des thèses*. Université de Genève, Faculté des sciences économiques et sociales.

Studer, M., Ritschard, G., Gabadinho, A., and Müller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods and Research*, 40(3):471–510.

van Driel, K. and Oosterveld, P. (2001). Nonoptimal alignment: A comment on "Measuring the agreement between sequences" by Dijkstra and Taris. *Sociological Methods & Research*, 29(4):524–531.

Widmer, E. and Ritschard, G. (2009). The de-standardization of the life course: Are men and women equal? *Advances in Life Course Research*, 14(1-2):28–39.

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 Journal of Sociology*, 29(1):35–65.

Wilson, C. (2006). Reliability of sequence-alignment analysis of social processes: Monte Carlo tests of ClustalG software. *Environment and Planning A*, 38:187–204.

Wu, L. L. (2000). Some comments on 'Sequence analysis and optimal matching methods

in sociology: Review and prospect'. *Sociological Methods Research*, 29(1):41–64.

Yujian, L. and Bo, L. (2007). A normalized Levenshtein distance metric. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 29(6):1091–1095.