# Harpoon or maggot?
# A comparison of various metrics to fish for sequence patterns

Nicolas Robette, Printemps (UVSQ-CNRS, UMR 8085)
Xavier Bry, I3M, Université de Montpellier II

## INTRODUCTION

### *Life course analysis : from causal to descriptive approaches*

Since the mid-1970s, life course analysis has become a major field of interest in the social sciences. It implies a shift "from structure to process, from macro to micro, from analysis to synthesis, from certainty to uncertainty" (Willekens, 1999, p.26, quoted in Ritschard & Oris, 2005). The development of life course analysis is simultaneously linked to theoretical issues as well as to advances in longitudinal micro-individual data collection and statistical analysis. The use of longitudinal data, such as panels or retrospective event-history surveys, has developed considerably (GRAB, 1999). On the methodological side, new analysis techniques have spread slowly but cumulatively, and social scientists now have a full toolbox. Since the early 1980s, the central approach has been *event history analysis* (Kalbfleisch & Prentice, 1980 ; Allison, 1984 ; Courgeau & Lelièvre, 1992 ; Mayer & Tuma, 1990). This set of techniques, such as the famous Cox model (1972), is a generalization of life table methods. Focusing on the occurrence of specific events, they model transition likelihoods or durations, under the assumption that life courses result from a complex stochastic process (Courgeau & Lelièvre, 1986). As a consequence, more often than not, event history analysis is a parametric approach, with a causal view.

Although most empirical life course studies rely on an event-based approach, theory has underlined the importance of the concept of trajectory (Sackmann & Wingens, 2003): events should not be studied independently from each other, but rather as a sequence. This implies a "'holistic' perspective that sees life courses as one meaningful conceptual unit" (Billari, 2001, p.440). A holistic approach allows to summarize the timing and sequence of events, durations in the various states and durations between events (Settersten & Mayer, 1997). Contrary to event history analysis, trajectory-based methods are mostly non-parametric: they make no assumption about the process underlying life courses and they belong to the "algorithmic model culture" (Breiman, 2001). They chiefly aim at exploring and describing life course as a whole and at "fishing for patterns" (Abbott, 2000). Event-based and trajectory-based approaches can thus be considered as two complementary cultures in life course analysis (Billari, 2005; Bry & Antoine, 2004): the latter intends to identify what differentiates life courses as a whole, while the former focuses on the risk of experiencing events (and its determinants).

### *Which metric should one choose[1] ?*

More often than not, the first step of holistic approaches consists in measuring the dissimilarity between life courses (regarded as sequences). Pairwise distances between sequences can further be used in various ways, often with data reduction techniques such

[1].Although we are referring here to life course analysis, which is the field where sequence analysis has expanded rapidly, the methods under study would also apply to any kind of sequence data, as do the results we present in the remainder of the article.

as multidimensional scaling or clustering. Many dissimilarity metrics exist in various domains (bioinformatics, data mining...) and their use in social sciences has been developing rapidly for a decade or two. The most widely known is certainly Optimal Matching Analysis (Abbott & Forrest, 1986), but other metrics for sequence analysis have been proposed and similar techniques using correspondence analysis also exist. Therefore, a crucial and pervasive issue in papers using holistic approaches is robustness: to what extent do the various techniques lead to consistent and converging results? What kinds of patterns does each of the metrics identify most effectively?

Numerous articles have been devoted to comparing metrics. However, most of them have limitations: they deal with a narrow range of methods at a time; they apply to specific sets of empirical data; other choices implied in the holistic approach (clustering techniques, etc.) may blur the results. So generalization is often problematic. We propose a systematic comparison of a collection of metrics that have been used in the social science literature, based on the examination of dissimilarity matrices computed from two data sets: a simulated one comprising various sequence patterns that sociologists may aim at identifying, and an empirical one (about occupational careers) as a "control sample". Thus what we are trying to do here is not to point out a hypothetical "best metric", but rather to unravel the specific patterns to which each alternative is actually more sensitive.

We will successively present a short review of existing methods for sequence analysis, a summary of the comparisons conducted in the literature, our own protocol for comparison, and finally, our results and discussion.

## VARIOUS DISSIMILARITY METRICS

The numerous dissimilarity metrics can be broadly grouped into two families: one linked to sequence analysis algorithms and the other to the tradition of geometric data analysis.

### *'Algorithmic' sequence analysis metrics*

Sequence analysis is a set of techniques for handling longitudinal data as ordered strings of elements. Among these techniques, Optimal Matching has been used, discussed and criticized far more often than the others. Optimal Matching was first developed in signal treatment and bioinformatics (Sankoff & Kruskal, 1983), and was introduced into the social sciences by Andrew Abbott during the 1980s (Abbott & Forrest, 1986). Its principle consists in measuring the dissimilarity between two sequences by transforming one sequence into the other. This can be done through three kinds of elementary operations: insertion (one element is inserted into a sequence), deletion (one element is deleted from a sequence) and substitution (one element is replaced by another in a sequence). Each operation can be assigned a cost. Thus the distance between two sequences is equal to the lowest cost needed to transform one sequence into the other.

To illustrate optimal matching, we will present two imaginary examples of sequences (Table 1). These sequences characterize school-to-work transitions, observed from ages 18 to 29 (i.e. made of 12 elements) and with 3 possible states: student (S); unemployed or inactive (U); employed (E). Calvin is at school up to the age of 19, then remains unemployed for three years and finally gets a job at age 23. Hobbes is a student until age 20, experiences two one-year spells of unemployment at ages 21 and 23, and works during the rest of the trajectory. These two sequences can be matched in different ways by means of insertions, deletions and substitutions. For instance, Hobbes' sequence can be aligned with Calvin's by deleting an S at the beginning of the sequence, inserting an E at the end and replacing E by U between the U spells: three operations are necessary. Another

possibility consists in replacing S by U at 20, E by U at 22 and U by E at 23: this also requires three operations.

*Table 1. Two examples of school-to-work transitions*

| *age* | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Calvin | S | S | U | U | U | E | E | E | E | E | E | E |
| Hobbes | S | S | S | U | E | U | E | E | E | E | E | E |

Cost setting of the elementary operations is a crucial step in optimal matching techniques: "The assignment of transformation costs haunts all optimal matching analyses" (Stovel *et al*, 1996). It underpins the method's flexibility and its ability to fit the object of research (Lesnard, 2010). Practically speaking, insertion and deletion are considered as a single operation, called *indel*, as deleting an element in one sequence is equivalent to inserting it in the other.

*Indel* operations pay more attention to the order of events in the sequences, by matching identical subsequences located at distinct positions within the sequences. The associated drawback is that they distort the timing of events. Substitutions, on the other hand, preserve the time structure by comparing simultaneous situations, but they alter the sequence of events. The balance between these costs will determine which kind of regularities are best captured in the dissimilarities computed.

Substitution costs are often chosen first. They may be constant whatever the states substituted, or distinct for each pair of possible states. In the latter case, costs can be driven by theoretical issues such as social stratification (Halpin & Chan, 1998; Blair-Loy, 1999; Scherer, 2001) or by the data themselves. Data-driven substitution costs are usually based on transition likelihoods between the states: the more frequent a transition, the more similar the states, the lower the costs. This cost scheme has become widely used in recent years (Han & Moen, 1999; Rohwer & Pötter, 2005; Aassve *et al*, 2007; Robette & Thibault, 2008). Other strategies can also be imagined, for instance combining a theoretical hierarchy of states and transition rates (Abbott & Hrycak, 1990; Stovel & Bolan, 2004).

Regarding *indel* costs, it is their relationship with substitution costs which is central. Some scholars choose equal substitution and *indel* costs, arguing that there would be no theoretical grounds for acting differently (Dijkstra & Taris, 1995). Thus, the distance between two sequences is equal to the smallest number of operations needed to match them. This is known as the *Levenshtein I distance* (Levenshtein, 1965). Moreover, optimal matching users often used to set high *indel* costs. However, with sequences of equal length, if *indel* costs are higher than the maximum substitution cost multiplied by half the sequence length, insertions and deletions will never be used. This is equivalent to the *Hamming distance* (Hamming, 1950), based on the simultaneousness of identical elements: the dissimilarity between two sequences is equal to the number of substitutions needed to match them, i.e. the number of time units where the situation is different. With sequences of different length, such a high *indel* cost leads to the use of insertions and deletions only to compensate for the length difference. Conversely, when *indel* costs are lower than half the minimum substitution cost, only *indel* operations will be used. The dissimilarity between two sequences thus corresponds to the length of their longest common subsequence (LLCS), which is also called *Levenshtein II distance*. In the end, cost setting comes down to positioning the cursor between Hamming and Levenshtein II

distances, depending on a preference for contemporaneousness of events or the existence of common subsequences (Lesnard & de Saint Pol, 2009).

While the use of optimal matching has increased significantly over the last decade, these techniques have been criticized (Levine, 2000; Wu, 2000; Elzinga, 2003)[2], giving birth to a "second wave" of sequence analysis (Aisenbrey & Fasang, 2010).

For instance, the choice of costs is often considered as arbitrary and weakly related to theoretical grounds, as elementary operations have no straightforward sociological interpretation (Levine, 2000; Wu, 2000; Elzinga, 2003). As a consequence, distances have no intrinsic meaning. Another criticism focuses on the order of events. Substitution costs are symmetrical, as replacing A by B in a sequence is equivalent to replacing B by A, and so the order of events is not properly taken into account (Wu, 2000). Moreover, optimal matching does not handle the direction of time (Wu, 2000). Indeed if transformation costs are identical at any point in time, then non-linear time dependency of sequences is neglected. With respect to the latter limit, Lesnard proposed a modified version of optimal matching, called *Dynamic Hamming Matching*, which uses time-varying costs (Lesnard, 2010). Practically, a substitution costs matrix is computed at each point in time, from the transition likelihoods between the various states at this particular moment, and insertions and deletions are prohibited. This variant has been applied in a few studies (Glorieux *et al*, 2008; Fasang, 2010; Lesnard & Kan, 2011).

Another set of sequence analysis methods, called *non alignment techniques*, has been developed. These metrics have the distinctive characteristic of not using the elementary operations needed in optimal matching. Their principle consists in calculating the similarity between sequences by comparing pairs of ordered elements of the sequences. DT coefficients (Dijkstra & Taris, 1995) compute the number of common pairs of ordered elements between two sequences. Elzinga's metrics can be considered as extensions of DT coefficients (Elzinga, 2003; 2006). They are also based on order relationships between pairs of elements. For instance, Elzinga suggested evaluating the longest common prefix between sequences (LCP), the length of the longest common subsequence (LLCS), the number of common subsequences (NCS) or the number of matching subsequences (NMS). The latter has been used successfully in some life course studies (Elzinga & Liefbroer, 2007; Bras *et al*, 2010; Liefbroer & Elzinga, 2012).

A few recent options for cost setting in optimal matching can also be reported (Gauthier *et al*, 2009 ; Hollister, 2009 ; Halpin, 2010). Lastly, another alternative has been proposed in a recent issue of this journal (Rousset *et al*, 2012): we refer the reader to the original article for a detailed presentation. It should be noted that the foundations of Rousset *et al*'s method, i.e. taking into account the fact that transition likelihoods may be unequally spread over the life course, are rather similar to those of Lesnard's *Dynamic Hamming Matching*.

*'Geometric data analysis tradition' metrics*

A second family of dissimilarity measures between sequences makes use of the broad range of geometric data analysis techniques. These metrics have seldom been used in the Anglo-Saxon social science literature (Van der Heijden, 1987; Martens, 1994; Van der Heijden *et al*, 1997). Yet numerous examples exist in French studies, particularly in the field of school-to work transitions and careers[3]. This might be related to the long-standing tradition of

[2] Most of these criticisms have been widely discussed in Abbott (2000).

[3] See for instance Degenne *et al* (1996), Béduwé *et al* (1995) or Robette & Thibault (2008). For a comprehensive review, see Grelet (2002) or Robette (2011).

correspondence analysis (or "geometrical analysis") in French statistics (Bry, 1995 et 1996 ; Lebart *et al*, 2000, Le Roux & Rouanet, 2004).
Several variants exist to apply geometric data analysis to life courses. The main difference between them lies in the way trajectories are coded and on the balance between chi-squared and Euclidean distance.
A first way of coding life courses consists in transforming them into an indicator matrix. In our example, 3 dummy variables are created for each year (one for each possible state): they are equal to 1 if the individual is in the given state during the given year, 0 otherwise. Thus there are 12*3=36 dummy variables. For instance, Calvin's trajectory would be coded as follows:

Table 2. Indicator matrix of Calvin's life course

| 18S | 18U | 18E | 19S | 19U | 19E | 20S | 20U | 20E | ... | 29S | 29U | 29E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | ... | 0 | 0 | 1 |

*Reading: At age 18, Calvin is a student, and not unemployed or in employment, and likewise at age 19 etc…*

After the coding step, the indicator matrix may used as input data for Correspondence Analysis (CA), which applies chi-squared distance. However, unscaled Principal Component Analysis (PCA) may be used instead of CA, which implies Euclidean distance instead of chi-squared (Espinasse, 1993; Béduwé *et al*, 1995). In the former case, the distance between two elements of the sequence is weighted by the inverse of the variable frequency: states which are infrequent in a given year contribute more to the measure of the dissimilarity between two life courses than the most frequent states. In other words, rare situations are given more importance. On the contrary, Euclidean distance makes the states' contribution equal and corresponds to the number of discordances between sequences. The final dissimilarity matrix is computed from the coordinates on the dimensions of CA or PCA.

Dissimilarity metrics using an indicator matrix and CA or PCA focus on the contemporaneousness of identical situations, whether these identical situations follow each other or are located at distant moments of the life course. The resemblance between trajectories is based upon the duration of simultaneousness in common states. The simultaneousness implies that the timing of situations or events is taken into account. On the other hand, the nature of transitions and their unfolding – in other words the sequence – are not part of the dissimilarity measure.

Another way of coding life courses can be viewed as a summary of indicator matrices. It is sometimes called *Qualitative Harmonic Analysis* (QHA) (Deville, 1974; Deville & Saporta, 1980; Dureau *et al*, 1994; Robette & Thibault, 2008). More precisely, the period under study is divided into sub-periods, then the time spent in each of the states for each sub-period is computed: this creates a number of variables equal to the number of sub-periods multiplied by the number of states. This set of variables is used as input for a Correspondence Analysis.

Coding the life course data for QHA involves several steps. First, the period investigated (here from 18 to 29 years old) is divided into sub-periods. Then, for each sub-period, the proportion of its duration spent in each state is computed. The number of variables created is equivalent to the number of sub-periods multiplied by the number of states (Table 3). Finally, these variables are used as input for Principal Component Analysis.

Table 3. Example of QHA coding of Calvin's life course

| S,18-21 | U,18-21 | E,18-21 | S,22-25 | U,22-25 | E,22-25 | S,26-29 | U,26-29 | E,26-29 |
|---|---|---|---|---|---|---|---|---|
| 0.5 | 0.5 | 0 | 0 | 0.25 | 0.75 | 0 | 0 | 1 |

*Reading : Calvin spent half of the first sub-period (from 18 to 21) as a student, the other half in unemployment, etc.*

A number of sub-periods equal to the period's length is equivalent to PCA with disjunctive coding. On the other hand, a unique sub-period would focus solely on durations in states. Therefore, the choice of the number of sub-periods is a trade-off between these two borderline cases. Moreover, the sub-periods do not have to be of equal duration. On the contrary, short sub-periods can be chosen for eventful years of the life course and longer sub-periods for quieter years. The ability to highlight "dense" moments of the life course (Rindfuss *et al*, 1987) is one of the major advantages of this metric.

Compared to the codings based on indicator matrices, the division into sub-periods makes the metric less sensitive to exact simultaneousness in common states. In other words, two sequences of states that are identical but slightly shifted will be considered as more similar by QHA metrics than by the two previous ones.

## A FEW EXISTING COMPARISONS

Although comparisons between methods are not a central issue for most authors[4], a (certain) number of papers test the influence of different alternatives while detailing their methodological protocol.

Some of them choose Optimal Matching and test various cost settings. For instance, drawing data from a study of the diffusion of Morris dancing in England, Forrest and Abbott (1990) introduce variation in substitution costs and conclude that the method seems to behave robustly. Reviewing this work in a further article, Abbott & Hrycak (1990) conclude that "the method thus seems to behave robustly with respect to variation […] in substitution costs […]. As is often the case, while care is needed, differences in minor analytic decisions are unlikely to drastically change results". Chan (1995) tests three substitution cost matrices on career data and notes "an impressive common core" across clusters and adds that the variation "follows interpretable patterns". In the same way, Anyadikes-Danes and McVicar perform several sensitivity analyses by introducing changes in substitution and *indel* costs in a study of school-to-work transitions. These highlight the fact that "relatively well-defined careers show up whether the cost matrix is designed to pick them out or not" (Anyadikes-Danes & McVicar, 2010) and that clusters are similar in nature, although there may be differences in cluster size and membership (McVicar & Anyadikes-Danes, 2002).

Other studies compare OMA with alternative metrics. Lesnard (2010) applies his own technique - i.e. dynamic Hamming matching - to time-use data and it turns out to be quite similar to Hamming distance, while Levenshtein II distance is a little less sensitive to contemporaneousness. Robette & Thibault (2008) study occupational careers with both Optimal Matching and Qualitative Harmonic Analysis. They note that the main career clusters are very similar whatever the technique and observe that OMA distinguishes

[4] Exceptions are Grelet (2002), Robette & Thibault (2008) or Aisenbrey & Fasang (2010).

slightly better between stable and mobile careers, while QHA seems slightly more sensitive to the presence of rare states. Aisenbrey & Fasang (2010) complete their review of sequence analysis methods by a comparative overview of dynamic Hamming matching, OMA with transition-based substitution costs and Elzinga's Number of Matching Subsequences (ignoring durations) applied to school-to-work transitions. The first two metrics lead to the same substantive patterns, despite mild differences in size and sensitivity to temporal variation. Still, NMS "departs more radically": it finds several "internally homogeneous clusters" and "one extremely heterogeneous clusters […] that comprises more than one half of all cases".

Finally, Grelet (2002) focuses on Geometric Data Analysis techniques and applies Principal Component Analysis, Correspondence Analysis and Qualitative Harmonic Analysis to school-to-work transitions. The results obtained largely converge, although CA and QHA seem more sensitive to rare situations than PCA.

On the whole, most of these comparisons broadly agree that pattern fishing remains robust, whatever the metric: "As is often the case, while care is needed, differences in minor analytic decisions are unlikely to drastically change results" (Abbott & Hrycak, 1990).

Another interesting result is that the variations observed in these empirical studies are consistent with the statistical groundings of the metrics. For example, CA and QHA use chi-squared distance, which weights dissimilarities by the inverse of the states' frequency on the sequences: thus it is not surprising that they attach more importance to rare states. In the same way, the use of *indel* operations in Optimal Matching shifts subsequences backward or forward, which logically makes Hamming distance (which utilizes no *indel* operations) less sensitive to sequence and transitions than Levenshtein II distance (which utilizes only *indel* operations).

However, these comparisons have several limitations. First, they focus on only a few metrics at a time: none of them tries to encompass a wide range of methods. Second, they all use one given set of empirical data. As a consequence, it is difficult to assess whether the conclusions remain valid beyond the specific case under study. And last but not least, comparisons are always performed by examining clusters of sequences. Yet building a typology of sequences implies a long string of methodological choices: choice of sequence analysis metric, but also choice of coding, clustering technique, parameterization and number of clusters. If the comparison is led from the end of the string, it is not easy to disentangle the mutual influence of the various steps of the analysis.

The strategy we adopt in the remainder of this paper attempts to overcome – to some extent – these limitations. To do so, we build an artificial set of sequences and then compare the dissimilarity matrices themselves to explore similarities and differences between a large range of sequence analysis metrics.

## COMPARING METRICS USING A SELECTED SET OF ARTIFICIAL SEQUENCES

### *A "reasoned" set of sequences*

Most of the papers comparing some of the numerous sequence analysis methods use the empirical data they aim to study. The drawback of this approach is that while the results may be interesting, the extent to which they can be generalized is questionable: it only solves the robustness issue in a very specific context. On the other hand, analyzing randomly simulated data would be pointless, as most of the time actual sequences don't look like random data at all, particularly for life course analysis. Even if they include a

certain amount of diversity, life courses usually have strong regularities and similarities. For these reasons, we chose to build a reasoned set of artificial sequences. This set is designed to contain the various kinds of regularities or differences that life courses analysts usually have to address: shifts, swaps, insertions, deletions, replacements, repetitions of spells (Barban & Billari, 2011), etc.

The sequences in our artificial set are of equal length (l=20), as some metrics do not handle length differences in a straightforward and unambiguous way. Possible states are the following 8 letters: A, B, C, D, E, F, G, H.

Practically, the set of sequences is divided into subsets. Each subset corresponds to a specific sequence of spells, with variable durations in these spells.

#1) *Time warping*: A subset of sequences A-B-C with varying durations in A, B and C (n=171)

#2) *Shifts*: Sequences A-B-C with B spell of fixed length equal to 6 and varying durations in A and C (n=13)

#3) *Reversal*: Initial sequences (subset #1) in reversed order, i.e. C-B-A (n=171)

#4) *Swaps*: Initial sequences (subset #1) with B and C swapped (i.e. A-C-B) or A and B swapped (i.e. B-A-C) (n=342)

#5) *Total permutation*: Initial sequences (subset #1) with all spells swapped, i.e. C-A-B and B-C-A (n=342)

#6) *Short insertion*: Sequence A-B-C with one short insertion (l=1) of spell D - i.e. D-A-B-C, A-D-B-C, A-B-D-C and A-B-C-D – with varying durations in A, B and C (n=612)

#7) *Long insertion*: Sequence A-B-C with one long insertion (l=10) of spell D - i.e. D-A-B-C, A-D-B-C, A-B-D-C and A-B-C-D – with varying durations in A, B and C (n=144)

#8) *Two shorts identical insertions*: Sequence A-B-C with two short insertions (l=1) of spell D - i.e. D-A-D-B-C, D-A-B-D-C, D-A-B-C-D, A-D-B-D-C, A-D-B-C-D, and A-B-D-C-D – with varying durations in A, B and C (n=821)

#9) *Two long identical insertions*: Sequence A-B-C with two long insertions (l=7) of spell D - i.e. D-A-D-B-C, D-A-B-D-C, D-A-B-C-D, A-D-B-D-C, A-D-B-C-D, and A-B-D-C-D – with varying durations in A, B and C (n=60)

#10) *Two shorts distinct insertions*: Sequence A-B-C with two short insertions (l=1) of spell D and E - i.e. D-A-E-B-C, D-A-B-E-C, D-A-B-C-E, A-D-B-E-C, A-D-B-C-E, and A-B-D-C-E – with varying durations in A, B and C (n=821)

#11) *Two long distinct insertions*: Sequence A-B-C with two long insertions (l=7) of spell D and E - i.e. D-A-E-B-C, D-A-B-E-C, D-A-B-C-E, A-D-B-E-C, A-D-B-C-E, and A-B-D-C-E – with varying durations in A, B and C (n=60)

#12) *One deletion*: Sequence A-B-C with A, B or C spell deleted, i.e. B-C, A-C (subset #12a) and A-B (subset #12b), with varying durations in A, B and C (n=57)

#13) *Two deletions*: Sequence A-B-C with A and B, B and C or A and C spells deleted, i.e. sequences A, B and C (n=3)

#14) *One replacement*: Initial sequences (subset #1) with A, B or C spell replaced by F, i.e. F-B-C, A-F-C and A-B-F (n=523)

#15) *Two replacements*: Initial sequences (subset #1) with A and B, B and C or A and C spells replaced by F and G spells, i.e. F-G-C, A-F-G and F-B-G (n=523)

#16) *Three replacements*: Initial sequences (subset #1) with A spell replaced by F, B by G and C by H, i.e. F-G-H (n=171)

#17) *Repetition of one spell*: Sequences A-B-A with varying durations in A and B spells (n=171)

#18) *Repetition of two spells*: Sequences A-B-A-B with varying durations in A and B spells (n=969)

#19) *Repetitions of a subsequence*: Sequences A-B-A-B-A-B-A-B-A-B-A-B-A-B-A-B-A-B-A-B-A-B-A-B (n=1)

#20) *Shifted repetitions of a subsequence*: Sequences B-A-B-A-B-A-B-A-B-A-B-A-B-A-B-A-B-A-B-A-B-A-B-A (n=1)

#21) *Repetition of the whole A-B-C sequence*: Sequences A-B-C-A-B-C with varying durations in A, B and C spells (n=11 628)

These subsets represent a total of 17 604 sequences. In order to reduce computation costs and to balance the weights of the various subsets, in subsets comprising more than 50 sequences, 50 sequences were selected at random. We end up with an artificial set of 854 sequences.

Moreover, in order to check some of our results against real life course data, we use a sample of 1 341 French male occupational careers drawn from the *Biographies et entourage* survey (INED, 2000). These data provide a record of occupations held from ages 14 to 50: the length of sequences is constant and equal to 37. There are 9 distinct states: farmer, self-employed, intermediate occupation, higher-level occupation, clerical and sales worker, manual worker, economically inactive, military conscript[5].

*The set of metrics*

Our aim is to provide comparisons of a comprehensive set of dissimilarity metrics[6].

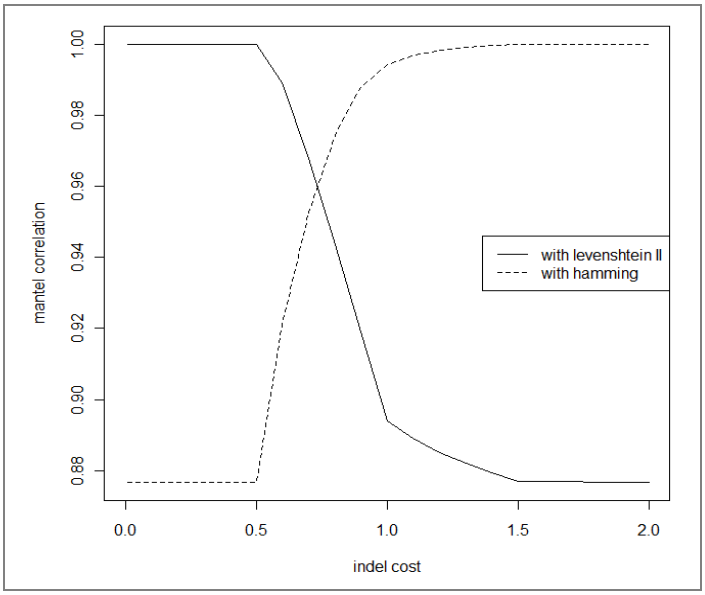
Regarding sequence-based metrics, we first use optimal matching with various cost schemes. The trade-off between substitution and *indel* costs determines which kinds of operations will be favoured. On the one hand, Hamming distance only uses substitutions, while on the other, Levenshtein II distance exclusively uses insertions and deletions. With sequences of equal length, both *indel* and substitution operations may be used if *indel* costs are somewhere between half the minimum substitution cost at one extreme and the maximum substitution cost multiplied by half the sequence length at the other. In order to uncover the optimal matching process between these two extremes, dissimilarity matrices are computed for the artificial set of sequences, with a constant substitution cost (s=2) and various *indel* costs. Cost schemes are then compared by computing correlation – by means

[5].For a detailed presentation of these data, see Robette & Thibault (2008).

[6] The following analyses have been computed with R, more specifically the TraMineR package (Gabadinho *et al*, 2011), and CHESA software for Number of Matching Subsequences (Elzinga, 2007).

of Mantel tests – between these matrices and dissimilarity matrices using Hamming distance on one side and Levenshtein II distance on the other (Figure 1).

Figure 1 – Mantel correlation between dissimilarity matrices with varying optimal matching cost schemes




First, our approach shows a Mantel correlation of 87.7% between Hamming and Levenshtein II metrics. As stated before, with an *indel* cost lower than 1, i.e. half the substitution cost, optimal matching is equivalent to Levenshtein II metric, which means that only *indel* operations are used. On the other side, equal *indel* and substitution costs, i.e. Levenshtein I distance, lead to a correlation of 99.4% with Hamming distance: they are almost equivalent. The drop in correlation occurs with an *indel* cost between 0.5 and 1: an *indel* equal to 0.75 times the substitution cost can be seen as a median setting.

The same analysis applies to our occupational career data leads to the same results, with even higher correlations: the Mantel correlation between Hamming and Levenshtein II reaches 97.3% (Appendix 1).

On the basis of these first results, the following 'algorithmic sequence analysis' metrics are chosen for further analyses: Hamming distance (HAM); Levenshtein II distance (LEVII); optimal matching with substitution costs based on transition likelihoods and a high *indel* cost, i.e. indel=2, in order to test the consequences of this common mode of substitution cost setting (OMAtr); Dynamic Hamming Distance (DHD); the variant proposed by Rousset *et al* (2012) (ROUS); Elzinga's Number of Matching Subsequences with durations handled through vector product (NMS). For metrics relating to geometric data analysis techniques, indicator matrix with Principal Component Analysis (PCA), indicator matrix with Correspondence Analysis (CA) and Qualitative Harmonic Analysis (QHA) are chosen[7].

Life course analysis can be defined as the statistical analysis of the timing of events or states, their sequencing, their quantum – i.e. the number of events or episodes – and the durations in the states (Billari, 2005). The metrics we have presented theoretically often take several of these dimensions into account. But in order to compare the extent to which

[7] This set of metrics does not claim to be exhaustive, but rather aims to take into account most of the most widely used sequence analysis techniques in social sciences.

they do so, it can be interesting to add metrics to our analysis that focus exclusively on one of them. Hamming distance focuses on the timing, through the simultaneousness in the states. For duration, we define a dissimilarity metric equivalent to the sum of squared differences in durations in the various states (DUR). For example, Calvin spends 2 years in S, 3 years in U and 7 years in E, while Hobbes spends 3 years in S, 2 years in U and 7 years in E: the dissimilarity between Calvin's and Hobbes' sequences is equal to $(2-3)^2+(3-2)^2+(7-7)^2=2$. Moreover, sequences can be represented in terms of episodes. Thus Calvin's sequence would take the shape S/2; U/3; E/7 and Hobbes' S/3;U/1;E/1;U/1;E/6, the figures indicating the number of consecutive years spent in the states. We can now define a dissimilarity metric focusing on quantum of episodes (QUA), equivalent to the sum of squared differences in the number of episodes in the states. For instance, Calvin spends 1 episode in each of the 3 states, while Hobbes spends 1 episode in S, 2 in U and 2 in E: the dissimilarity between their sequences is equal to $(1-1)^2+(1-2)^2+(1-2)^2=2$. Last, concerning sequence of episodes, we use the Length of the Longest Common Subsequence metric ignoring duration (SEQ), i.e. Levenshtein II distance on sequences represented as S-U-E and S-U-E-U-E (respectively for Calvin and Hobbes), for example.

We finally have a set of 12 metrics: the 9 examined here and 3 "control" metrics (DUR, QUA and SEQ).

## RESULTS

### *Three sets of similar metrics*

Each of these metrics is then used to compute a dissimilarity matrix between the sequences of the "reasoned" set: we thus have 12 dissimilarity matrices. Next the matrices are compared by measuring Mantel correlation between each pair of matrices (Table 4).

*Table 4. Mantel correlations between dissimilarity matrices, using different metrics*

| | DUR | QUA | SEQ | LEVII | HAM | OMAtr | DHD | ROUS | PCA | CA | QHA | NMS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DUR | 100 | 34,9 | 34,3 | 88,9 | 72,6 | 75,1 | 73,4 | 72,1 | 70,1 | 61,5 | 62,7 | -1,8 |
| QUA | 34,9 | 100 | 82,4 | 37,5 | 28,4 | 30,6 | 30,3 | 27,8 | 20,1 | 28,8 | 29,3 | 67,5 |
| SEQ | 34,3 | 82,4 | 100 | 53,2 | 46,6 | 49,1 | 49,3 | 45,0 | 36,8 | 45,8 | 45,9 | 52,1 |
| LEVII | 88,9 | 37,5 | 53,2 | 100 | 87,7 | 90,4 | 89,2 | 86,5 | 83,4 | 75,1 | 75,8 | -1,2 |
| HAM | 72,6 | 28,4 | 46,6 | 87,7 | 100 | 99,3 | 99,7 | 99,2 | 96,9 | 72,4 | 72,0 | -0,6 |
| OMAtr | 75,1 | 30,6 | 49,1 | 90,4 | 99,3 | 100 | 99,6 | 98,2 | 95,4 | 75,7 | 75,5 | -1,0 |
| DHD | 73,4 | 30,3 | 49,3 | 89,2 | 99,7 | 99,6 | 100 | 98,6 | 95,8 | 75,5 | 75,2 | -0,8 |
| ROUS | 72,1 | 27,8 | 45,0 | 86,5 | 99,2 | 98,2 | 98,6 | 100 | 97,9 | 70,7 | 70,2 | -0,3 |
| PCA | 70,1 | 20,1 | 36,8 | 83,4 | 96,9 | 95,4 | 95,8 | 97,9 | 100 | 63,1 | 62,5 | -6,2 |
| CA | 61,5 | 28,8 | 45,8 | 75,1 | 72,4 | 75,7 | 75,5 | 70,7 | 63,1 | 100 | 99,6 | -3,8 |
| QHA | 62,7 | 29,3 | 45,9 | 75,8 | 72,0 | 75,5 | 75,2 | 70,2 | 62,5 | 99,6 | 100 | -3,7 |
| NMS | -1,8 | 67,5 | 52,1 | -1,2 | -0,6 | -1,0 | -0,8 | -0,3 | -6,2 | -3,8 | -3,7 | 100 |

First of all, we notice that the metrics are closer to the duration measure (DUR) – correlations range from 61% to 89% – than to the sequence measure (SEQ) – from 45% to 53% - and above all the quantum measure (QUA) – from 20% to 38% –, except Elzinga's metric (NMS). Indeed the latter has a null correlation with the duration measure but a relatively strong correlation with measures based on quantum (67%) and sequence (52%). Among the others, LEVII seems to better combine these dimensions of temporality: it has the highest correlations with DUR, QUA and SEQ.

More generally, looking at the whole set of correlations, three groups of metrics may be distinguished: a first group that we will call "OM-like" metrics, which comprises LEVII, HAM, OMAtr, DHD, ROUS and PCA; a small group of "CA-like" metrics (CA and QHA); and NMS.

In the first group, correlations range from 83% to almost 100%. Among others, we see that OMAtr, DHD and ROUS's correlations with Hamming distance (HAM) are above 99%: they are almost equivalent. PCA is very close as well, as its correlation with these other four metrics ranges from 95% to 98%: the emphasis on contemporaneousness brings them together. LEVII, through the use *indel* operations, gives less priority to contemporaneousness, its correlation with the other "OM-like" metrics nonetheless remains high (from 83% to 91%).

In the second group, CA and QHA are almost equivalent (the correlation is 99.6%). Their correlation with "OM-like" metrics is high: between 62% and 76%.

NMS, conversely, is totally orthogonal to the other metrics: its correlation is almost null with any of them.

Most of these results still hold when the same approach is applied to actual sequence data, i.e. male occupational career data from the "*Biographies et entourage*" survey[8]. The major differences are an even higher homogeneity of "OM-like" metrics (for instance, LEVII and HAM have a correlation of 97.3%) and a lower correlation between "CA-like" and "OM-like" metrics (from 45% to 57%).

*Different ways of fishing for patterns*

Now we have a global picture of the relative resemblances between metrics, we aim to uncover what drives the differences between them, in other words to determine the kind of sequence patterns to which each of these metrics is more sensitive.

For each metric, the distances between sequences of the "reasoned" set are computed, the distances are ranked[9] and then the ranks are scaled to make them comparable (the scaled rank is equal to 0 if the sequences are considered identical, and the most distant sequences have a scaled rank of 100).

We then compare the results of the different metrics for the specific sequence patterns one may want to fish (Table 5). For example, if the focus is on time-warping, scaled ranked distances between sequences from the first subset will be studied, i.e. ABC sequences with varying durations in A, B and C. If now the focus is on reversals, we will be interested in scaled ranked distances between sequences from the first subset and sequences from the third subset, i.e. between ABC and CBA sequences).

[8] Although our standpoint consists in comparing directly the distances produced by the metrics rather than the typologies of sequences, which imply additional methodological choices, we have tested the robustness of our results with typologies. We have thus compared typologies (in 5, 10, 15 and 20 clusters) obtained from each of the metrics, and from our two data set, by measuring the resemblance between typologies using the Jaccard index. The 4*2=8 comparison matrices lead to the same conclusions about proximities between metrics (tables available from the authors).

[9] Indeed, we are not interested in the distance level itself, but rather in the relative distances, i.e. the fact that two given sequences will be considered as more similar than two other sequences or not. Moreover, for some metrics the distance distribution is highly skewed, while it is not so for others: scaled distances would not be appropriate for a comparison between metrics.

Table 5. Scaled ranked distances between sequences, for different metrics and sequence patterns

| Patterns | DUR | QUA | SEQ | LEVII | HAM | OMAtr | DHD | ROUS | PCA | CA | QHA | NMS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| time warping (#1 vs #1) | 20 | 0 | 0 | 14 | 12 | 13 | 13 | 13 | 18 | 11 | 11 | 2 |
| shifts (#2 vs #2) | 10 | 0 | 0 | 6 | 14 | 15 | 15 | 15 | 19 | 11 | 11 | 1 |
| reversal, ie ABC vs CBA (#1 vs #3) | 20 | 0 | 50 | 44 | 50 | 55 | 55 | 55 | 64 | 43 | 45 | 23 |
| swaps, ie ABC vs ACB or BAC (#1 vs #4) | 20 | 0 | 10 | 24 | 25 | 28 | 27 | 28 | 33 | 21 | 21 | 8 |
| total permutation, ie ABC vs CAB or BCA (#1 vs #5) | 20 | 0 | 10 | 31 | 52 | 51 | 55 | 56 | 63 | 39 | 39 | 15 |
| 1 insertion of a short D spell (#1 vs #6) | 22 | 12 | 4 | 15 | 14 | 15 | 15 | 15 | 18 | 11 | 12 | 9 |
| 1 insertion of a long D spell (#1 vs #7) | 44 | 12 | 4 | 35 | 35 | 39 | 37 | 38 | 41 | 44 | 43 | 29 |
| 2 insertions of short D spells (#1 vs #8) | 24 | 27 | 10 | 17 | 16 | 17 | 17 | 17 | 18 | 11 | 12 | 55 |
| 2 insertions of long D spells (#1 vs #9) | 61 | 27 | 10 | 55 | 52 | 56 | 54 | 56 | 61 | 54 | 54 | 68 |
| 2 insertions of short D and E spells (#1 vs #10) | 26 | 27 | 10 | 18 | 18 | 19 | 19 | 19 | 20 | 14 | 15 | 52 |
| 2 insertions of long D and E spells (#1 vs #11) | 61 | 27 | 10 | 55 | 52 | 56 | 55 | 56 | 60 | 73 | 72 | 60 |
| 1 deletion, ie ABC vs AB (#1 vs #12b) | 33 | 12 | 4 | 26 | 24 | 26 | 25 | 26 | 34 | 19 | 20 | 5 |
| 1 deletion, ie ABC vs AC or BC (#1 vs #12a) | 34 | 12 | 4 | 27 | 21 | 23 | 23 | 23 | 29 | 19 | 20 | 6 |
| 2 deletions, ie ABC vs A, B or C (#1 vs #13) | 56 | 27 | 10 | 50 | 34 | 38 | 37 | 37 | 49 | 32 | 34 | 7 |
| 1 replacement, ie ABC vs ABF, AFC or FBC (#1 vs #14) | 42 | 27 | 10 | 35 | 27 | 31 | 31 | 29 | 31 | 46 | 49 | 17 |
| 2 replacements, ie ABC vs AFG, FBG or FGC (#1 vs #15) | 69 | 65 | 50 | 64 | 49 | 57 | 56 | 53 | 51 | 69 | 71 | 28 |
| 3 replacements, ie ABC vs FGH (#1 vs #16) | 90 | 87 | 84 | 90 | 82 | 95 | 94 | 90 | 91 | 90 | 90 | 31 |
| AB vs ABA (#12b vs #17) | 19 | 12 | 4 | 19 | 18 | 18 | 18 | 20 | 26 | 20 | 21 | 5 |
| AB vs ABAB (#12b vs #18) | 17 | 27 | 10 | 14 | 11 | 11 | 11 | 12 | 15 | 10 | 11 | 20 |
| many repetitions of AB spells (#12b vs #19) | 11 | 100 | 100 | 19 | 11 | 12 | 12 | 13 | 6 | 6 | 8 | 100 |
| slight shift of "ABABABABABABABABABAB" (#19 vs #20) | 0 | 0 | 10 | 0 | 82 | 1 | 79 | 84 | 0 | 0 | 0 | 100 |
| overall repetition, ie ABC vs ABCABC (#1 vs #21) | 20 | 52 | 30 | 18 | 20 | 21 | 22 | 22 | 24 | 16 | 13 | 91 |

This approach gives a similar picture as in the previous section: there are still three distinct groups of metrics, “OM-like” (LEVII, HAM, OMAtr, DHD, ROUS and PCA), “CA-like” (CA and QHA) and NMS.

Compared to “OM-like” metrics, “CA-like” metrics seem more sensitive to insertions of one long spell or two long different spells, and to one or two replacements. Conversely, they are less sensitive to short insertions. This means that “CA-like” metrics more easily capture differences in the universe of states composing sequences, insofar as the states appearing in one sequence and not in the other correspond to long spells. Moreover, these metrics are a little less sensitive to time warping and shifts, reversals, swaps, total permutations and repetitions: they attach less importance to the way and the order in which spells unfold. Lastly, they consider two highly unstable and slightly shifted sequences (subset #19 vs subset #20) as totally similar.

As one might expect from the previous stages of our inquiry, NMS behaves noticeably differently from “OM-like” metrics. It is highly sensitive to repetitions of spells. It is also significantly more sensitive to two insertions, especially when short spells are inserted, i.e. when the sequence of spells composing the whole sequence differs, even if the differing spells have a short duration. Furthermore, NMS is less sensitive to differences in duration, i.e. to time warping and shifts, but above all to reversals, swaps, total permutations, deletions and replacements. This appears harder to interpret: it seems that NMS's focus on sequence of spells operate only in specific cases, in particular when “alien” spells are short. More strangely, NMS considers ABC and FGH sequences (i.e. subset #1 vs subset #16) as rather similar, although they do not share any common state. On the other hand,

two highly unstable and only slightly shifted sequences (subset #19 vs subset #20) are viewed as totally distinct.

Let us go a little further by comparing metrics from the "OM-like" group. Compared to Hamming distance (HAM), PCA is somewhat more sensitive to time warping and shifts, reversals, swaps and total permutations, deletions and long insertions. ROUS, OMAtr and DHD are almost equivalent to HAM, except that the two latter and, to a lesser extent, the former, capture replacements a bit more easily. Unsurprisingly, because of the use of *indel* operations which gives less importance to contemporaneousness, Levenshtein II distance (LEVII) is less sensitive then HAM to shifts and more notably to total permutations. Moreover, it captures deletions and replacements better. But the main difference among this group regards a special case, that of the comparison between two unstable and slightly shifted sequences (subset #19 vs subset #20): while LEVII, OMAtr and PCA ignore the shift, HAM, DHD and ROUS consider the two sequences as highly distinct.

As far as "CA-like" metrics are concerned, CA and QHA do not differ in a significant way for any kind of pattern.

## CONCLUSION

Since the 1980s, sequence analysis approaches have become widespread in the social sciences, and Optimal Matching has been the leading method. But OMA has been discussed and amended, and other metrics have been proposed. So there is a crucial need for comparisons between existing metrics. They are based on different statistical traditions (algorithmic culture, geometric data analysis, etc.) and all have specificities, particularly in the way they handle the various dimensions of temporality, e.g. contemporaneousness, durations or order. We have proposed an approach, based on a "reasoned" set of sequences, to uncover what kind of patterns each sequence analysis method is more able to fish for.

The results confirm what was already suspected: social science sequence data are strongly structured, in such a way that the main patterns they conceal will be uncovered by most of the metrics. But as marginal differences may be of importance, it is useful to understand precisely the kinds of sequences to which these differences are tied. We have revealed three groups of heavily converging metrics - 1) Levenshtein II, Hamming and Dynamic Hamming distances, OMA with data-driven substitution costs, Rousset et *al.*'s metric and Principal Component Analysis; 2) Correspondence Analysis and Qualitative Harmonic Analysis; 3) Elzinga's Number of Matching Subsequences – as well as the small distinctions among them. This constitutes a further step towards a better knowledge of the wide range of available sequence analysis methods, so that scholars can pick the one that best suits their data design and inquiry objectives.

## References


Aassve Arnstein, Billari Francesco C., Piccarreta Raffaella, 2007, “Strings of adulthood: a sequence analysis of young british women's work-family trajectories”, *European Journal of Population,* 23(3-4), p. 369-388.

Abbott Andrew, 2000, “Reply to Levine and Wu”, *Sociological methods & research,* 29(1), p. 65-76.

Abbott Andrew, Forrest John, 1986, “Optimal Matching Methods for Historical Sequences”, *Journal of Interdisciplinary History,* 16(3), p. 471-494.

Abbott Andrew, Hrycak Alexandra, 1990, “Measuring ressemblance in sequence data : an optimal matching analysis of musicians’ careers”, *American journal of sociology,* (96), p. 144-185.

Aisenbrey Silke, Fasang Anette E., 2010, “New Life for Old Ideas: The ‘‘Second Wave’’ of Sequence Analysis Bringing the ‘‘Course’’ Back Into the Life Course”, *Sociological Methods & Research,* 38(3), p. 420-462.

Allison Paul David, 1984, *Event history analysis: regression for longitudinal event data*, Beverly Hills, CA, Sage (coll. Quantitative applications in the social sciences), vol. 46, 87 p.

Anyadike-Danes Michael, McVicar Duncan, 2010, ““My Brilliant Career: Characterizing the Early Labor Market Trajectories of British Women From Generation X”, *Sociological Methods & Research,* 38(3), p. 482-512.

Barban Nicolas, Billari Francesco, 2011, “Classifying life course trajectories: a comparison of latent class and sequence analysis”, *Dondena Working Papers*, 41, 27 p.

Béduwé Catherine, Dauty Françoise, Espinasse Jean-Michel, 1995, Trajectoires types d'insertion professionnelle. Application au cas des bacheliers professionnels de Midi-Pyrénées, in *Deuxièmes journées d'étude Céreq-Lasmas-IdL "L'analyse longitudinale du marché du travail", 28 et 29 juin 1995*, Caen, Céreq, p. 7-29.

Billari Francesco C., 2001, “Sequence analysis in demographic research”, *Canadian Studies in Population,* 28(2), p. 439-458.

Billari Francesco C., 2005, “Life course analysis: two (complementary) cultures? Some reflections with examples from the analysis of the transition to adulthood”, *Advances in life course research,* 10, p. 261-281.

Bison Ivan, 2009, ““OM matters: the interaction effects between *indel* and substitution costs”, *Methodological Innovations Online*, 4(2), p. 53-67.

Blair-Loy Mary, 1999, “Career patterns of executive women in finance: an optimal matching analysis”, *The American Journal of Sociology,* 104(5), p. 1346-1397.

Bras Hilde, Liefbroer Aart C., Elzinga Cees H., 2010, “Standardization of Pathways to Adulthood? An Analysis of Dutch Cohorts Born Between 1850 and 1900”, *Demography*, 47(4), p. 1013-1034.

Breiman Leo, 2001, “Statistical Modeling: The Two Cultures”, *Statistical Science,* 16(3), p. 199-231.

Bry Xavier, 1995, *Analyses factorielles simples*, Paris, Economica (coll. Techniques quantitatives - poches), 112 p.

Bry Xavier, 1996, *Analyses factorielles multiples*, Paris, Economica (coll. Techniques quantitatives - poches), 112 p.

Bry Xavier, Antoine Philippe, 2004, “Exploring explanatory models. An event history application”, *Population-E,* 59(6), p. 795-830

Chan Tak Wing, 1995, “Optimal matching analysis: a methodological note on studying career mobility”, *Work and occupations,* 22(4), p. 467-490.

Courgeau Daniel, Lelièvre Eva, 1986, “Nuptialité et agriculture”, *Population,* (2), p. 303-326.

Courgeau Daniel, Lelièvre Eva, 1992, *Event history analysis in demography*, Clarendon Press, Oxford, 226 p.

Cox D. R., 1972, “Regression models and life tables (with discussion)”, *Journal of royal statistical society,* (B34), p. 187-220.

Degenne Alain, Lebeaux Marie-Odile, Mounier Lise, 1996, Typologies d’itinéraires comme instrument d’analyse du marché du travail, in Degenne Alain, Mansuy Michèle, Podevin Gérard, Werquin Patrick, (eds) *Typologie des marchés du travail, suivi et parcours, 23 et 24 mai 1996*, Rennes, (coll. Documents séminaire Céreq), vol. 115, p. 27-42.

Deville Jean-Claude, 1974, “Méthodes statistiques et numériques de l’analyse harmonique”, *Annales de l'INSEE,* (15), p. 3-101.

Deville Jean-Claude, Saporta Gilbert, 1980, “Analyse harmonique qualitative”, in Diday Edwin (éds)*, Data analysis and informatics*, Amsterdam, North Holland Publishing, p. 375-389.

Dijkstra W, Taris T, 1995, “Measuring the agreement between sequences”, *Sociological methods & research,* (24), p. 214-231.

Dureau Françoise, Barbary Olivier, Elisa Flores C., Hoyos M. C., 1994, La observacion de las diferentes formas de movilidad : propuestas metodologicas experimentadas en la encuesta de movilidad espacial en el area metropolitana de Bogota, in *Atelier du CEDE (Montevideo), Nuevas modalidades y tendencias de la migracion entre paises fronterizos y los processos de integracion, 27-29 octobre 1993*, Paris, Ortsom, p. 31.

Elzinga Cees H., 2003, “Sequence similarity: a nonaligning technique”, *Sociological methods & research,* 32, p. 3-29.

Elzinga Cees H., 2006, “Sequence analysis: metric representations of categorical time series”, *Sociological methods & research*, under revision

Elzinga, Cees H., 2007, *CHESA 2.1 User Manual*, Amsterdam: Vrije Universiteit Amsterdam.

(Available from home.fsw.vu.nl/ch.elzinga/.)

Elzinga Cees H., Liefbroer Aart C., 2007, "De-standardization of family-life trajectories of young adults: a cross-national comparison using sequence analysis", *European Journal of Population,* 23(3-4), p. 225-250.

Espinasse Jean-Michel, 1993, "Enquêtes de cheminement, chronogrammes et classification automatique", *Note du Lhire,* 19(159).

Fasang Anette E., 2010, "Retirement: Institutional Pathways and Individual Trajectories in Britain and Germany", *Sociological Research Online*, 15(2), 16 p.

Forrest John, Abbott Andrew, 1990, "The optimal matching method for studying anthropological sequence data", *Journal of Quantitative Anthropology,* 2, p. 151-170.

Gabadinho Alexis, Ritschard Gilbert, Müller Nicolas, Studer Matthias, 2011, "Analyzing and visualizing state sequences in R with TraMineR", *Journal of Statistical Software*, 40(4), p. 1-37.

Gauthier Jacques-Antoine, Widmer Éric D., Bucher Philipp, Notredame Cédric, 2009, "How Much Does It Cost?: Optimization of Costs in Sequence Analysis of Social Science Data", *Sociological Methods & Research,* 38(1), p. 197-231.

Glorieux Ignace, Mestdag Inge, Minnen Joeri, 2008, "The Coming of the 24-hour Economy?: Changing work schedules in Belgium between 1966 and 1999", *Time & Society*, 17(1), p. 63-83.

GRAB, 1999, *Biographies d'enquêtes : bilan de 14 collectes biographiques*, Paris, INED (coll. Méthodes et savoirs), vol. 3, 340 p. *[http://grab.site.ined.fr/fr/editions_en_ligne/biographies_enquetes/]*

Grelet Yvette, 2002, "Des typologies de parcours. Méthodes et usages", *Document Génération 92,* (20), 47 p.

Halpin Brendan, 2010, "Optimal Matching Analysis and Life-Course Data: The Importance of Duration", *Sociological Methods & Research,* 38(3), p. 365-388.

Halpin Brendan, Chan Tak Wing, 1998, "Class careers as sequences: an optimal matching analysis of work-life histories", *European Sociological Review,* 14(2), p. 111-130.

Hamming R.W., 1950, "Error-detecting and error-correcting codes", *Bell System Technical Journal,* 29(2), p. 147-160.

Han Shin-Kap, Moen Phyllis, 1999, "Clocking out: temporal patterning of retirement", *American journal of sociology,* 105(1), p. 191-236.

Hollister Matissa, 2009, "Is Optimal Matching Suboptimal? ", *Sociological Methods & Research,* 38(2), p. 235-264.

Kalbfleisch John D., Prentice Robert L., 1980, *The statistical analysis of failure time data*, New York, Wiley (coll. Wiley series in probability and mathematical statistics), 322 p.

Lebart Ludovic, Morineau Alain, Piron Marie, 2000, *Statistique exploratoire multidimensionnelle*, Paris, Dunod (coll. Sciences sup), 439 p.

Le Roux Brigitte, Rouanet Henri, 2004, *Geometric Data Analysis: From Correspondence Analysis to Stuctured Data Analysis*, Kluwer Academic Publishers, Dordrecht, 475 p.

Lesnard Laurent, 2010, "Setting Cost in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns", *Sociological Methods & Research,* 38(3), p. 389-419.

Lesnard Laurent, Kan Man Yee, 2011, "Investigating scheduling of work: a two-stage optimal matching analysis of workdays and workweeks", *Journal of royal statistical society (series A)*, 174(2), p. 349-368.

Lesnard Laurent, de Saint Pol Thibaut, 2009, "Décrire des données séquentielles en sciences sociales : panorama des méthodes existantes", communication aux *Xe Journées de Méthodologie Statistique*, 23-25 mars 2009, Paris, INSEE.

Levenshtein V.I., 1966[1965], "Binary codes capable of correcting deletions, insertions, and reversals", *Soviet Physics Doklady,* 10, p. 707-710.

Levine Joel H., 2000, "But what have you done for us lately? Commentary on Abbott and Tsay", *Sociological methods & research,* 29(1), p. 34-40.

Liefbroer Aart C., Elzinga Cees H., 2010, "Intergenerational Transmission of Behavioural Patterns: How Similar are Parents' and Children's Demographic Trajectories?", *Advances in Life Course Research*.

Martens Bernd, 1994, "Analyzing event history data by cluster analysis and multiple correspondence analysis: an example using data about work and occupations of scientists and engieneers", in Greenacre Michael, Blasius Jorg *Correspondence analysis in the social sciences: recent developments and applications*, New-York, p. 233-251.

Mayer Karl Ulrich, Tuma Nancy Brandon, 1990, *Event history analysis in life course research*, Madison, University of Wisconsin Press, 297 p.

McVicar Duncan, Anyadike-Danes Michael, 2002, "Predicting successful and unsuccessful transitions from school to work by using sequence methods", *Journal of royal statistical society A,* (165), p. 317-334.

Rindfuss Ronald R., Swicegood C. Gray, Rosenfeld Rachel A., 1987, "Disorder in the life course: how common and does it matter ?", *American sociological review,* 52(6), p. 785-801.

Ritschard Gilbert, Oris Michel, 2005, "Life course data in demography and social sciences: statistical and data-mining approaches", *Advances in life course research,* 10, p. 283-314.

Robette Nicolas, 2011, *Explorer et décrire les parcours de vie: les typologies de trajectoires*, CEPED (Les Clefs pour...), 86 p.

Robette Nicolas, Thibault Nicolas, 2008, "Comparing Qualitative Harmonic Analysis and

Optimal Matching. An Exploratory Study of Occupational Trajectories", *Population-E,* 63(4), p. 533-556.

Rohwer Götz, Pötter Ulrich, 2005, *TDA's user manual*, 1021 p.

ROUSSET ET AL, 2012, BMS

Sackmann Reinhold, Wingens Matthias, 2003, "From transitions to trajectories: Sequence types", in Heinz Walter R., Marshall Victor W. *The Life Course: Sequences, Institutions and Interrelations*, New-York, Aldine de Gruyter, p. 93-112.

Sankoff David, Kruskal Joseph, (dir), 1983, *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*, Reading, Addison-Wesley, 408 p.

Scherer Stefani, 2001, "Early career patterns: a comparison of Great Britain and West Germany", *European Sociological Review,* 17(2), p. 119-144.

Settersten Jr Richard A., Mayer Karl Ulrich, 1997, "The measurement of age, age structuring and the life course", *Annual review of sociology,* 23, p. 233-261.

Stovel Katherine, Bolan Marc, 2004, "Residential trajectories. Using optimal alignment to reveal the structure of residential mobility", *Sociological methods & research,* 32(4), p. 559-598.

Stovel Katherine, Savage Michael, Bearman Peter, 1996, "Ascription into achievement: models of career systems at Lloyds Bank, 1890-1970", *American Journal of Sociology,* 102(2), p. 358-399.

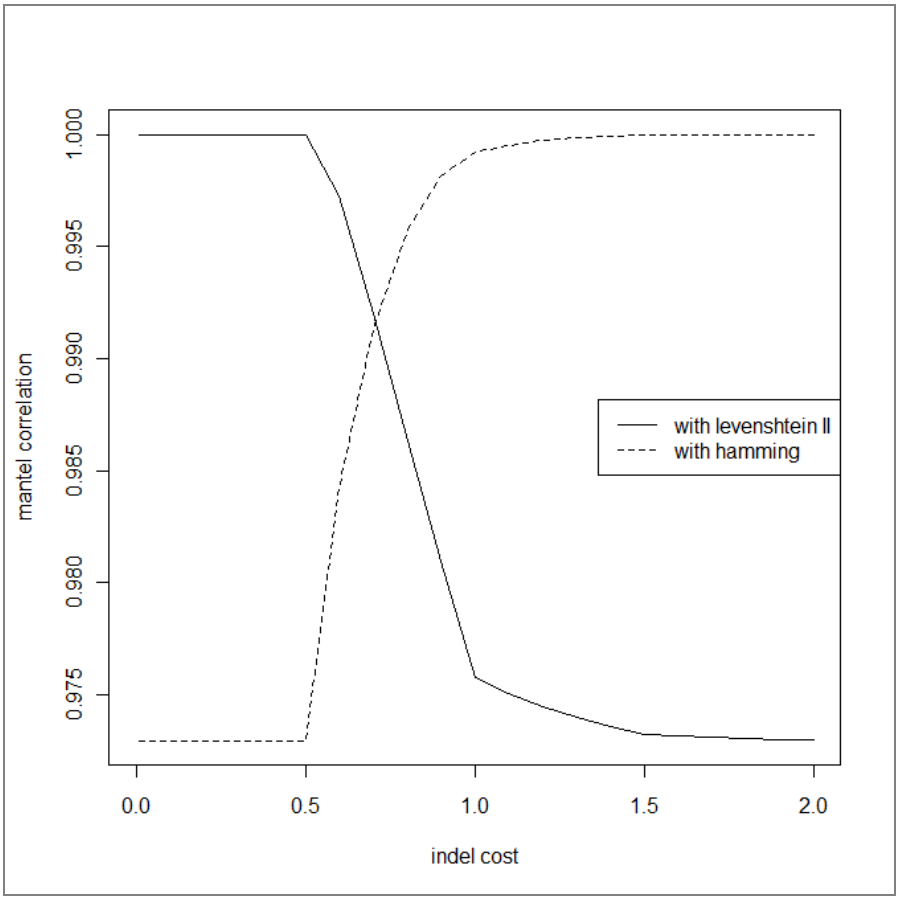
Van der Heijden Peter G. M., 1987, *Correspondence analysis of longitudinal categorical data*, Leiden, DSWO Press.

Van der Heijden Peter G. M., Teunissen Joop, van Orlé Charles, 1997, "Multiple correspondence analysis as as tool for quantification or classification of career data", *Journal of Educational and Behavioral Statistics,* 22(4), p. 447-477.

Willekens F.J., 1999, "The life course: Models and analysis", in van Wissen L.J.G., Dykstra P. (eds) *Population issues: An interdisciplinary focus*, New-York, Plenum Press, p. 23-51.

Wu Lawrence L., 2000, "Some comments on "Sequence analysis and optimal matching methods in sociology: Review and prospect"", *Sociological methods & research,* 29(1), p. 41-64.

APPENDIX 1 - Mantel correlation between dissimilarity matrices with varying optimal matching cost schemes, using occupational career data



APPENDIX 2 - Mantel correlation between dissimilarity matrices with different metrics, using occupational career data

| | DUR | QUA | SEQ | LEVII | HAM | OMAtr | DHD | ROUS | PCA | CA | AHQ | NMS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DUR | 100 | 34,9 | 34,3 | 88,9 | 72,6 | 75,1 | 73,4 | 72,1 | 70,1 | 61,5 | 62,7 | -1,8 |
| QUA | 34,9 | 100 | 82,4 | 37,5 | 28,4 | 30,6 | 30,3 | 27,8 | 20,1 | 28,8 | 29,3 | 67,5 |
| SEQ | 34,3 | 82,4 | 100 | 52,4 | 54,5 | 54,6 | 54,5 | 54,1 | 44,1 | 42,2 | 40,8 | 36,6 |
| LEVII | 88,9 | 37,5 | 52,4 | 100 | 97,3 | 97,6 | 97,4 | 95,6 | 95,4 | 54,5 | 53,6 | 1,3 |
| HAM | 72,6 | 28,4 | 54,5 | 97,3 | 100 | 99,9 | 100,0 | 98,9 | 97,1 | 56,6 | 54,9 | 4,0 |
| OMAtr | 75,1 | 30,6 | 54,6 | 97,6 | 99,9 | 100 | 99,9 | 98,7 | 97,1 | 56,6 | 55,1 | 3,7 |
| DHD | 73,4 | 30,3 | 54,5 | 97,4 | 100,0 | 99,9 | 100 | 98,8 | 97,1 | 56,5 | 54,9 | 3,8 |
| ROUS | 72,1 | 27,8 | 54,1 | 95,6 | 98,9 | 98,7 | 98,8 | 100 | 97,7 | 56,7 | 54,6 | 4,9 |
| PCA | 70,1 | 20,1 | 44,1 | 95,4 | 97,1 | 97,1 | 97,1 | 97,7 | 100 | 46,8 | 45,4 | -1,8 |
| CA | 61,5 | 28,8 | 42,2 | 54,5 | 56,6 | 56,6 | 56,5 | 56,7 | 46,8 | 100 | 93,2 | 6,6 |
| AHQ | 62,7 | 29,3 | 40,8 | 53,6 | 54,9 | 55,1 | 54,9 | 54,6 | 45,4 | 93,2 | 100 | 4,0 |
| NMS | -1,8 | 67,5 | 36,6 | 1,3 | 4,0 | 3,7 | 3,8 | 4,9 | -1,8 | 6,6 | 4,0 | 100 |