METHODOLOGICAL INNOVATIONS Online

Methodological Innovations Online 4(2) 53-67

OM matters: the interaction effects between *indel* and substitution costs

Ivano Bison*

*Department of Sociology and Social Research, University of Trento

Abstract

There is increasing interest in, and use of, Optimal Matching techniques in sequence analysis. In recent years especially, both the uses and the software tools that enable this type of analysis have multiplied. Taking up Abbott's inheritance, many scholars have continued with epistemological and methodological analysis of the application of OM to the social sciences. Nevertheless, many criticisms and doubts have been raised concerning the capacity of the technique to capture common patterns of sequences. Diverse epistemological, methodological and technical questions are awaiting answers. Among the numerous problems one, which has not yet been noticed, concerns the interaction effects between OM elementary operations. This article discusses the interaction effect between *indel* and substitution costs in Optimal Matching Analysis (henceforth OMA). By means of a simulation based on the eight sequences obtained as element permutation of a binary string of length 3, it will show that varying the substitution and *indel* costs produces inconsistent results. The article will also show that the sequence patterns obtained from analysis of the distances matrix do not depend on the real existence of common patterns of sequences; rather, they are the result of the interaction effects between the elementary operations of substitution and *indel*.

Key words: Optimal matching; sequence analysis; simulation

Introduction

The idea of sequence in the study of social processes is not new. As early as 1855, Herbert Spencer discussed time and sequences at length in his *The Principles of Psychology*. Some years later, Durkheim (1912) spoke of *social rhythms*, and finally, George Herbert Mead forcefully asserted: "We look for their antecedents in the past and judge the future by the relation of this past to what is taking place" (Mead, 1932, p.45).

Nevertheless, it is only with the brilliant and persevering work of Andrew Abbott that the study of the social process by sequences has attracted the attention of sociologists. In 1983 Abbott began publication of a series of articles (Abbott, 1983, 1984, 1988, 1990, 1991a, 1991b, 1992, 1995, 2000; Abbott and Forrest, 1986; Abbott and Hrycak, 1990; Abbott and DeViney, 1992; Abbott and Tsay, 2000; MacIndoe and Abbott 2004) in which he laid the theoretical and methodological bases for Sequence Analysis. Abbott's intent with sequence analysis was not only to emphasize study of the social process in its becoming, instant after instant, searching

Correspondence: I. Bison Department of Sociology and Social Research, University of Trento, Via Verdi, 26, 38100 – Trento, Italy. E-mail: ivano.bison@soc.unitn.it

ISSN 1748-0612online
DOI:

for regularities, but also to find a new way to observe social reality which extended beyond the consolidated tradition of the linear-regression family (Abbott, 1988).

The purpose of sequence analysis is to give answers to some simple questions. How do observed social regularities come about? How does this pattern come about? How does the career process allocate people to different positions in the class structure in a way that ultimately produces the observed mobility regime? (Chan, 1999). In other words, sequence analysis is used to "fish for patterns" (Abbott, 2000, p.69) to account for the complexity of processes.

Abbott's problem was finding a rigorous way to give concrete application to sequence analysis: that is, to find a method able to discriminate between events pertaining to different rhythms and events with close cadence, thus identifying different sub-rhythms in social processes. Abbott found the solution in an algorithm used in biology and other sciences (Sankoff and Kruskal, 1983, Waterman, 1995) known as the "optimal matching method" (Abbott and Forrest, 1986). Optimal Matching is a family algorithm that takes account of the full complexity of sequence data. It is based on a dissimilarity measure of distance between sequences which was first proposed in the field of computer science by Vladimir Levenshtein (1966) and developed by, among others, molecular biologists studying protein or DNA sequences.

Whereas on the one hand, adoption of this technique enabled Abbott to argue more forcefully for the validity of his approach, on the other, the epistemological, methodological and technical implications connected with the use of OM in the social sciences raised doubts and criticisms (Dijkstra and Taris, 1995; Wu, 2000; Levine, 2000; Eltzinga, 2003). A further effect was that sequence analysis was increasingly identified with the OM algorithm, rather than adherence to the original idea. Studies therefore began to refer to sequence analysis as 'optimal matching analysis', and some scholars (e.g. Lesnard, 2006; Halpin, 2008) sought to develop a new epistemological and methodological reasoning which started from the method that operates the OM algorithm.

Despite the perplexities and criticisms, the OM technique progressively spread among researchers, also because of the diffusion of software tools – Optimize[1], TDA[2] (Rhower and Potter, 2009), Stata (Brzinsky-Fay *et al*. 2006) – enabling the conduct of OM analysis. Many prestigious journals began to publish studies which used the OM techniques (e.g. Chan, 1995; Stovel *et al*., 1996; Halpin and Chan, 1998; Han and Moen, 1999; Blair-Loy, 1999; Scherer, 2001; McVicar and Anyadike-Danes, 2002; Clark *et al*., 2003; Malo and Muñoz-Bullón, 2003; Stovel and Bolan, 2004; Anyadike-Danes and McVicar, 2005; Wilson, 2006; Levy *et al*. 2006; Lesnard, 2006; Pollock, 2007; Wiggins *et al*., 2007; Aasave *et al*., 2007; Martin *et al*., 2008).
The current widespread use of OM in the academic world might suggest that many problems have been resolved and that the technique has consolidated. But this is not the case. All studies, which use OMA, devote at least a section to defending and justifying it. Yet the criticisms have indubitably played a role in inducing researchers to find solutions to OMA problems. Not all the problems have been resolved, however, and doubts still persist as to its validity.

This article will seek to answer the following simple questions. Is it possible that the OM sequence distances do not represent the system under analysis? Is it possible that the generation of distances between different sequences produces distortions in representation of the system? And is it possible that, with variance in operations costs, the system configuration varies not according to the costs but according to the interaction effects among OM operators, thus producing distortions in the representation of the reality under examination?

[1] Optimize is freeware developed by Andrew Abbott: see http://home.uchicago.edu/~aabbott/om.html#optimize.
[2] TDA 6.4 is a freeware developed by Götz Rohwer and Ulrich Pötter: see http://steinhaus.stat.ruhr-uni-bochum.de/.

A series of simulations were conducted in order to answer these questions. The aim was to verify how the results of the analyses varied according to variation in substitution and *indel* costs: that is, whether the OM distances between sequences reflected the system configuration under examination, or whether it was only the result of the algorithm computation.

After briefly describing how OM operates, the principal criticisms and the OM operator cost definition, the article presents the results of the simulations. The analyses reported focused both on the standard method, which studies the entire configuration of the sequences in search of common patterns, and on the most recent proposal named "Ideal Type" (Scherer, 2001; Wiggins *et al*., 2007; Martin *et al*., 2008), which focuses on the deviations from a specific sequence (Abbott and Hrycak, 1990).

The Optimal Matching distance

In OM analysis, the distance between two sequences A and B is given by the number of elementary operations that enable sequence A to be turned into the sequence B, or *vice versa*. Two types of elementary operation can be performed to obtain this transformation. The first operation, named *indel*, inserts and deletes sequence elements. For instance, in order to transform sequence A {111} into sequence B {110} we must: a) move the 1 in the third position of sequence A one position to the right {11_1}; b) insert a 0 in the third position of sequence A, so as to get a new sequence A' {1101}; c) delete the 1 in the fourth position of A' {110~~1~~}. The result is sequence A" {110}, which is equal to sequence B. This operation is maximally useful when we have sequences of different lengths. For instance, let A {111} and C {11} be two sequences respectively of length 3 and 2. We can transform one into the other either by deleting the last element of A or by inserting a 1 at the end of C.

The second operation is substitution of the elements of one sequence with the elements of the other sequence. Sequence A in the first example can be converted into sequence B by replacing the third element of sequence A, the 1 with the 0; or sequence B can be converted into A by replacing the third element of B, the 0 with the 1. This operation is performed between elements occupying the same position in the two sequences. Its optimal use is when both the sequences have the same length.

Both elementary operations may be necessary to transform one sequence in another. Let A {111} and D {10} be two sequences with different lengths and elements. One will be transformed into the other through substitution of the second element (of A or D) and through insertion/deletion of the third element (of A or D). The situation becomes complicated when the *indel* and substitution operations can both be performed on the same element of a sequence. For instance, to transform sequence A into B we can either insert/delete or substitute the last element. What is to be done? Which of the two operators applies? The solution is to assign a "cost" to each operator. In other words, to the operators are assigned a quantity that represents the energy dissipated by the system in performing the operations of *indel* and substitution. In social terms, we can conceive the cost as the effort (economic, social, cultural, physical) that a subject expends to pass from one state to another: for instance, from being employed to unemployed, or from being married to being divorced.

The choice of which of the two operations to perform is determined by the smaller of the two costs, i.e. we choose the operation that minimizes the energy required to transform one sequence into another. If *indel* > substitution, we will substitute the element; if *indel* < substitution, we will insert/delete the element.

In this scenario, the OM is an algorithm that finds, for each element in the sequence, the appropriate operation which minimizes the cost of transformation into the corresponding element in the other sequence. The OM distance is instead the amount of all the operations that minimize the transformation cost of the entire sequence. That is, the OM distance is the least quantity of energy necessary to transform one sequence into another.

Limitations and criticisms of Optimal Matching Analysis

Many doubts and criticisms have been expressed in regard to OMA (Dijkstra and Taris, 1995; Wu, 2000; Levine, 2000; Eltzinga, 2003). The principal criticisms concern two separate aspects. The first is the epistemological and methodological validity of the use of OM operations to treat events in time. The second, more technical, aspect is the reliability of finding similarity between sequences.

Manipulating sequences by operators to assess their similarity is *de facto* to manipulate time. Inserting or deleting an event is also to warp the timing of the processes, whereas substituting an event means that the timing is preserved but that one event is approximated by another. As Lesnard writes: "… insertion and deletion operations preserve the events but distort time while substitution operations just do the opposite, i.e. they conserve time but alter events. As a result, OM with sequences of social events is a combination of accelerations/decelerations to match identical subsequences of events and of events approximations when the flow of time is normal" (Lesnard, 2006, p.7).

The second group of criticisms centre on four main problems. The first is the impossibility of defining a distance between sequences according to their degree of similarity. Even if two sequences have some elements in common but in different positions, they may be as dissimilar as a third sequence with no element in common with the other two. For instance, consider the following three different sequences {01; 10; 22} and a substitution and *idel* cost equal to one. The OM distance between all of these sequences will be equal to two.

OM distance: {01,10} = {10,22} = {01,22} = 2

This problem is not solved with an appropriate substitution cost matrix. Instead, matters become more complicated, and this is the second main problem. Serious doubts have been raised concerning the arbitrariness of the definition of the substitution cost matrix (Wu, 2000). Apart from the choice of the costs, a third problem concerns the assumption of symmetry of distance between two states (Wu, 2000). By definition, the substitution cost matrix is symmetric: hence, for instance, the cost of passing from an employment to an unemployment event is the same as the cost of moving from an unemployment to an employment event. This may produce considerable logical contradictions when non-reversing and non-consecutive conditions (such as never married, divorced) are to be replaced, or when the transition has very different social costs, as in the above example of movement between employed and unemployed.

The fourth problem is the impossibility of capturing both the inner timing of each single sequence and the general timing of the entire system. This depends on how the algorithm works. Comparison between two sequences is from left to right, starting from the first pair until the last pair. The algorithm starts from the first pair of elements to the left and decides what operations to apply. After finding the lowest cost operation, the algorithm passes to the following pair of elements and repeats the operation until the end of the two sequence elements. In this process, the algorithm does not consider the entire sequence but only the single pairs of elements. The result of this way of proceeding is that the algorithm is blind to the arrow of time "[in fact]…one obtains the same distance between trajectories if the trajectories move "forward" or "backward" in time" (Wu, 2000, p.52), because the choice of which operation to apply is independent of the temporal position along the sequence. The result is also memory loss, since the choice of operation is independent from the previous choice. Finally, "the substitution cost is blind to the environments of the pair of elements being considered" (Halpin, 2008, p.6). Consequently, not considering the whole sequence is not to recognize the possible presence of recursive structure.

The choice of cost system

The central point of the OM is the choice of a cost system that determines how sequences are matched. When only *indel* operations are used, or when all substitution costs are strictly greater than the cost of *indel*, then the distance between two sequences is equivalent to their longest common subsequence, whatever their location in the two sequences (Kruskal, 1983, p. 30). "On the contrary, using only substitution operations or when their cost is strictly lower than the cost of *indel* will focus the analysis on finding contemporaneous similarities" (Lesnard, 2006, p.4). In some way, "*indel* costs should be defined as a function of the temporal proximity of identically coded events, substitution costs should represent the closeness of two different events at a particular position in their respective sequences" (Lesnard, 2006, p.9). Because the costs do not have a measurable physical quantity, they are defined *a priori* on the quality/property of transition or transformation that the researcher assumes to exist between the elementary states that make up the sequences being analyzed.
In effect, there is no clear or univocal way to derive the substitution costs matrix. Martin *et al*. (2008) report three different strategies to determine substitution costs. Some analysts derive these costs from theory. For example, Halpin and Chan (1998) derive the substitution cost from the EGP class strata, while McVicar and Anyadike-Danes (2002) derive the costs between different employment states from the educational level required to access them, and in another case from "the degree of attachment to the labor market of the different activities" (Anyadike-Danes and McVicar 2005, p.515). Other scholars (Pollock *et al*., 2002; Pollock, 2007; Stovel *et al*., 1996) derive the costs as the inverse of the transition probability between two states. In this case, the cost-distance will be larger as smaller is the probability of observing transitions between the two states. Yet other authors prefer to fix all the costs of the substitution costs matrix to a specific value. For example, Martin *et al*. (2008) fix a substitution cost at one for all the states transitions.

As a rule, the *indel* operation has only one cost value, while the substitution operation has one or more costs for all the n(n-1)/2 couples of elements defined in the state space. Studies in the literature have focused mainly on the rules to define the substitution costs matrix. They have paid less attention to the rules to define the value of *indel*. The most frequently cited study is that by Abbott and Hrycak (1990, p.155), who suggest setting the value of *indel* "equaling or slightly exceeding the highest cost of substitution". This means that when sequences are of roughly equal length, most transformations will likely rely on substitution rather than insertion or deletion. Although the literature comprises numerous proposals for rules and ways to define the costs, *de facto* each article suggests and uses its own rules. There are studies which define the value of *indel* as equal to the maximum value of substitution costs (Stovel *et al*., 1996). Others make the *indel* costs vary according to the length of the sequences (Stovel and Bolan, 2004). Others set the *indel* costs below the maximum value of the substitution costs (Halpin and Chan, 1998). Others define the *indel* costs as greater than the maximum value of the substitution costs (McVicar and Anyadike-Danes, 2002). Others prefer to leave the default software parameter used for the OM calculation: usually *indel* cost = 1 and substitution costs = 2.
Yet few researchers have addressed the problem of the effects of the interaction between substitution and *indel* costs on the estimation of the distances between sequences. I believe that this is because costs have been defined mainly as a distance among states rather than establishing which of the two elementary operations the OM algorithm must use.

The simulation

Now described are the results of a simulation which studied the interaction effects on the computation of sequence distances when the substitution costs were made to vary and the *indel* cost was kept fixed. The simulation was based on the set of eight sequences 2^3 obtained by permutation of a binary string of length 3. Hence, the main characteristic of this dataset was that it contained the entire state space of all the possible sequences that can be realized with a binary sequence of length 3. The 0 and the 1 of the sequence can be interpreted as the coding of a realization of two mutually exclusive states, such as, for instance, being employed and not employed; married and not married; student and non-student. The Needleman-Wunsch Algorithm, implemented in the Stata program (Brzinsky-Fay *et al.*, 2006), was used for the distance matrixes/vectors calculations.

With the value of *indel* fixed at 1, the OM distances were computed for the following costs of substitution: 0.5, 0.6, 0.8, 1.0, 1.5, 2.0. Alternatively, these costs can be interpreted as proportions of the rate ‘SI’ between substitution and *indel* cost:

$$SI = \frac{Substitution\ cost}{Indel\ cost}$$

For instance, substitution and *indel* costs of respectively {0.8; 1.0}, {1.2; 1.5}, {1.6; 2.0} have exactly the same proportion rate SI = 0.8. This means that the cost value changes but the costs rate and the structure of the distance matrix/vector remain the same.

Two results were expected from the simulation:

a) The first was that the distances between sequences would vary in an equal and/or proportional way according to the variation in substitution costs.

b) The second was that varying substitution costs would not change the sequence groups with the same distance. As well known, one shortcoming of the OM it is that it produces the same distances for different sequences. The expectation was that groups of sequences with the same distance would remain stable with variance in the substitution costs.

Optimal Matching and Ideal Type

I begin with discussion of the OM distance calculated as deviation from a single reference sequence *ideal type*. For brevity, I shall discuss only the results of two analyses. The first took the sequence {001} as reference; and the second took the sequence {010} as reference. Graphically, the two sequences are similar; both have only one event (1) and two events (0). In qualitative terms, they partly describe different careers. For instance: a) the first sequence describes a work career where the subject is employed for the first two months of observation and unemployed in the last month of observation; b) the second sequence describes a work career where a subject is employed in the first and third month and unemployed in the second month of observation.

Table 1. OM distance vectors by 001 and 010 reference sequences by substitution cost equal to 0.6, 0.8, 1.0, 1.5, 2.0 and *indel* equal to 1.

	Reference: {001}					Reference: {010}				
Sequences	0.6	0.8	1.0	1.5	2.0	0.6	0.8	1.0	1.5	2.0
000	0.6	0.8	1.0	1.5	2.0	0.6	0.8	1.0	1.5	2.0
001	-	-	-	-	-	1.2	1.6	2.0	2.0	2.0
010	1.2	1.6	2.0	2.0	2.0	-	-	-	-	-
100	0.6	0.8	1.0	1.5	2.0	0.6	0.8	1.0	1.5	2.0
011	1.2	1.6	2.0	2.0	2.0	1.2	1.6	2.0	2.0	2.0
101	0.6	0.8	1.0	1.5	2.0	1.8	2.0	2.0	2.0	2.0
110	1.8	2.4	3.0	3.5	4.0	0.6	0.8	1.0	1.5	2.0
111	1.2	1.6	2.0	3.0	4.0	1.2	1.6	2.0	3.0	4.0
N. of clusters	3.0	3.0	3.0	4.0	2.0	3.0	3.0	2.0	3.0	2.0
SI	0.6	0.8	1.0	1.5	2.0	0.6	0.8	1.0	1.5	2.0
min	0.6	0.8	1.0	1.5	2.0	0.6	0.8	1.0	1.5	2.0
max	1.8	2.4	3.0	3.5	4.0	1.8	2.0	2.0	3.0	4.0
Rate (Max/Min)	3.0	3.0	3.0	2.3	2.0	3.0	2.5	2.0	2.0	2.0

The outcomes from the reference sequence {001} are substantially different from those expected (Table 1.). Only the sequences {000, 100, 101, 111} have a variation of the distance from the reference sequence equal and/or proportional to the variation of the substitution costs. The first three sequences {000, 100, 101} increase (Table 1.) the distance from the reference sequence in the same way as raising the substitution costs (0.6, 0.8, 1.0, 1.5, 2.0) does. The sequence {111} instead has a distance value double that of the substitution costs from the reference sequence (1.2, 1.6, 2.0, 3.0, 4.0). The last three sequences the distance raise equal and/or proportional with substitution costs less than 1.0 and not proportional for values greater than 1.0. The sequence {110} has a distance three times greater than the substitution cost for values up to 1.0; 2.3 times higher for a cost of 1.5; and double for a cost of 2.0. The sequences {010 and 011} have a distance value which is double for substitution costs greater than 1.0, 1.3 times higher for a substitution cost of 1.5, and equal to the substitution cost[3] for a value of 2.0.

The direct consequence of non-proportional variation in distances according to the variation in the substitution cost is that both the sequences with the same distance and number of clusters change. In the example (Table 2.), for costs up to 1.0 three distinct sequence groups are produced with the same distance; for a cost of 1.5 the groups increase to 4, and for a cost of 2.0 they decrease to 2.

Table 2. Groups sequences according to the distance from the reference sequence {001} by substitution cost.

0.6 = (000, 100, 101); (010, 011, 111); (110).
0.8 = (000, 100, 101); (010, 011, 111); (110).
1.0 = (000, 100, 101); (010, 011, 111); (110).
1.5 = (000, 100, 101); (010, 011); (111); (110).
2.0 = (000, 100, 101, 010, 011); (110, 111).

It is easy to imagine the effects of using these vectors of distance as input on the outcome of a cluster analysis: as substitution costs vary so will the number of clusters. The consequence will be that it is not possible to determine whether a specific solution is the result of a specific configuration of observed sequences or rather the result of the choice of a specific substitution cost. However, what happens if we take the sequence {010} as reference? The expectation is to find the same problems as previously arose with the sequence {001}. But, unfortunately, this is not so (Table 1.). Only the sequences {000, 100, 110} have distances equal to the

[3] In this last case, the distance from the reference sequence is constant to 2 for values of the substitution costs greater than 1.0.

substitution costs, while both the range of the distances and the numerousness of groups change for the other sequences. The sequences {001, 011} have double distances for substitution costs less than 1.0 and decreasing values for costs greater than or equal to 1.0. The sequence {101} has a three times greater distance for a substitution cost of 0.6 and decreasing ones for higher substitution costs.

In the previous analysis for costs up to 1.0, the clusters of sequences were the same. This may suggest that defining substitution costs less than or equal to 1.0 can be expected to yield a certain stability (at least for simple systems) in the number of sequence clusters. Unfortunately, this conclusion is refuted on changing the reference sequence (Table 3.). Now, the three clusters appear, but for different substitution costs: 0.6, 0.8 and 1.5. The four cluster configurations disappear, and two new cluster configurations appear. The first, for a substitution cost of 1.0, is made up of the sequences {000, 100, 110} and the sequences {001, 011, 101, 111}. The second configuration, for a substitution cost of 2.0, which has the same distance (2.0) for all the sequences except the sequence {111}.

Table 3. Groups sequences according to the distance from the reference sequence {010} by substitution cost

0.6= (000, 100, 110); (001, 011, 111); (101).
0.8= (000, 100, 110); (001, 011, 111); (101).
1.0= (000, 100, 110); (001, 011, 101, 111).
1.5= (000, 100, 110); (001, 011, 101); (111).
2.0= (000, 100, 110, 001, 011, 101); (111).

The problem is further complicated when the system's complexity increases to comprise several states and costs simultaneously. For instance, let us suppose that we have a sequence of length 3 with three states {employed, unemployed, not working} coded with {0, 1, 2}. Let us also suppose that we use the substitution matrix costs of Table 4. What happens when we take the sequence {001} or the sequence {002} as reference? In theory nothing should change. All substitution costs being equal (Table 4.), the distances from the reference sequence should be the same.

Table 4. Substitution costs matrix

	0	1	2
0	0.0		
1	0.6	0.0	
2	0.8	0.0	0.0

This invariance is also desired. In fact, in qualitative terms the two reference sequences are very similar. Both have two consecutive employment events (t1 and t2) in common, and they differ only in the last state t3: one to unemployment and one to not working. Nevertheless, this difference is very small. The distance between the two states of exit from work is zero (Table 4.), while that between work and unemployment and non-work events is respectively 0.6 and 0.8. Unfortunately, the result (Tables 5a. and 5b.) is different from expected.

Table 5a. Groups sequences according to the distance from the sequence {001}

Distance	0.0	0.6	0.8	1.2	1.4	1.6	1.8	2.2

Ref.Seq. {001}	001, 002	***000***, 011, 101	022, 202	010, 100, ***111***	020, 200	***222***	110	220

Table 5b. Groups sequences according to the distance from the sequence {002}

Distance	0.0	0.6	0.8	1.2	1.4	1.6	2.0	2.4
Ref.Seq. {002}	001, 002	011, 101	***000***, 022, 202	***111***	010, 100	020, 200, ***222***	110	220

Although the number of groups distance does not change, there is nevertheless a change in the composition of clusters according to the reference sequence used. The sequences {000, 111, 222} jump from one group to the other. This is due to the different substitution cost linked to the specific state defined in the reference sequence. That is, with different states in the same position, the results groups of sequences vary with variance in the substitution costs linked to the states defined in the reference sequence.

Optimal Matching and standard sequence analysis

The next step is to evaluate the effects of variations in costs on a standard OM sequence analysis. Here the distances calculation is not a single distance/vector from a specific reference sequence but a distance matrix between all the sequences under analysis. Table 6. shows two triangular distance matrixes: the first (lower triangle) obtained for a substitution cost of 2.0, the second (upper triangle) for a substitution cost of 0.8. The two matrixes are clearly different. For a substitution cost of 2.0 we can count only three distinct values of distance: one 6, eight 4s and nineteen 2s, while for a cost of 0.8 we can count four different values of distance: three 2.4s, one 2, twelve 1.6s and twelve 0.8s. What has happened? The problems previously found for the *ideal types* now concern all the sequences that are taken one by one as references.

Table 6. Optimal Matching distance matrix with *indel* = 1.0 and substitution cost matrix = 2 (lower triangle) and Optimal Matching distance matrix with *indel* = 1.0 and substitution cost matrix = 0.8 (upper triangle)

	000	001	010	100	011	101	110	111
000	-	0.8	0.8	1.6	0.8	1.6	1.6	2.4
001	2.0	-	1.6	0.8	1.6	0.8	2.4	1.6
010	2.0	2.0	-	0.8	1.6	2.0	0.8	1.6
100	4.0	2.0	2.0	-	2.4	1.6	1.6	0.8
011	2.0	2.0	2.0	4.0	-	0.8	0.8	1.6
101	4.0	2.0	2.0	2.0	2.0	-	1.6	0.8
110	4.0	4.0	2.0	2.0	2.0	2.0	-	0.8
111	6.0	4.0	4.0	2.0	4.0	2.0	2.0	-

If we analyze the opposite diagonal (Table 7.) of the distance matrix with variance in substitution costs – in other words, if we analyze the couples of sequences with maximum distances between them, that is, those sequences that do not have common elements and are complementary {000 – 111; 001 – 110; etc.} – we find that the values of the distances change with variance in substitution costs.

Table 7. Opposite diagonal of OM distance matrix by substitution costs

	0.6	0.8	1.0	1.5	2.0

000 – 111	1.8	2.4	3.0	4.5	6.0
001 – 110	1.8	2.4	3.0	3.5	4.0
010 – 101	1.8	2.0	2.0	2.0	2.0
100 – 011	1.8	2.4	3.0	3.5	4.0
011 – 100	1.8	2.4	3.0	3.5	4.0
101 – 010	1.8	2.0	2.0	2.0	2.0
110 – 001	1.8	2.4	3.0	3.5	4.0
111 – 000	1.8	2.4	3.0	4.5	6.0

For a cost of 0.6, all the couples have the same distance and values three times higher than the substitution costs. For values of 0.8 and 1.0, the distance is three times higher than the substitution costs, except for the couple {010 – 101}, which is 2.5 and 2.0 times the substitution cost. The costs of 1.5 and 2.0 have three values of distance. The first is three times higher than the substitution cost for the couple {000 – 111}. The second is 2.3 and 2.0 times the substitution cost for the couples {001 – 110 and 100 – 011}. Finally, the couple {010 – 101} has a distance equal to the substitution cost when the value is 2.0 and 1.3 times higher for a substitution cost of 1.5.

The empirical evidence clearly shows that the distance matrix changes with variance in substitution costs. However, it may be objected that these differences are irrelevant, and that the analyses on each single distance matrix can give exactly the same results. In other words, even if the distances changes, the configuration and the relationship between sequences do not change.

Following the literature, I analyzed the distance matrixes first with a multidimensional scaling (MDS) and then with a hierarchical cluster. “This matrix itself must then be analyzed, typically with some form of dual-data reduction scheme such as cluster analysis or multidimensional scaling.” (Abbott and Tsay, 2000, p.6). I began with the MDS. Given the small number of points, I considered it sufficient to produce a two-dimensional solution. I performed four separate analyses for the following four substitution costs distance matrixes: 0.5, 0.6, 1.0 and 2.0.

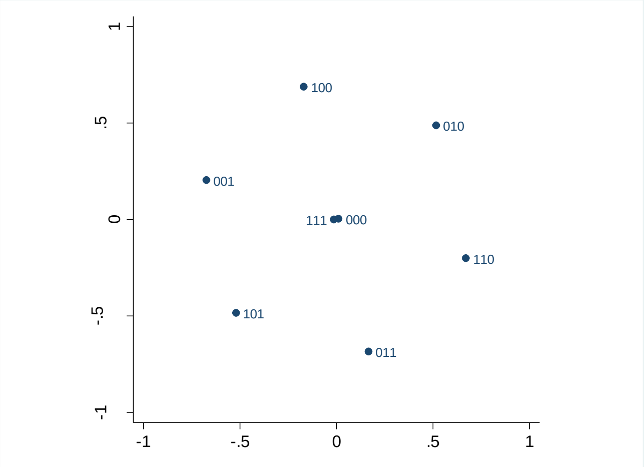


Fig.1. MDS *INDEL*=1.0; SUB=0.5; Fit(1)=0.50; Fit(2)=0.60

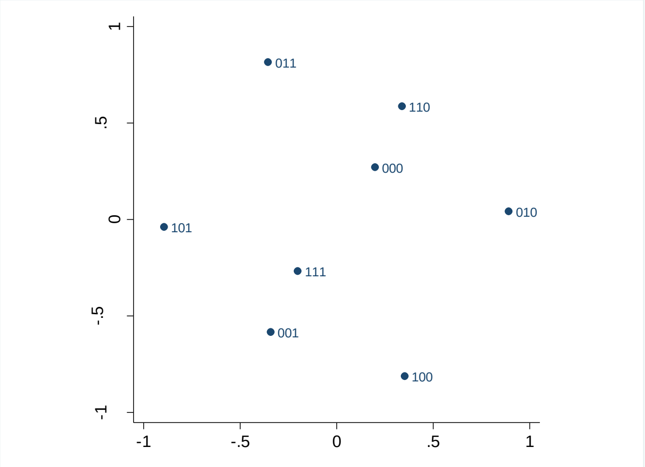
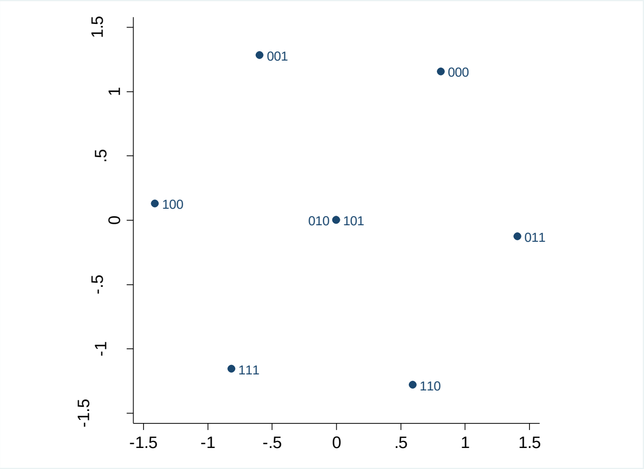


Fig.2. MDS *INDEL*=1.0; SUB=0.6; Fit(1)=0.50; Fit(2)=0.60



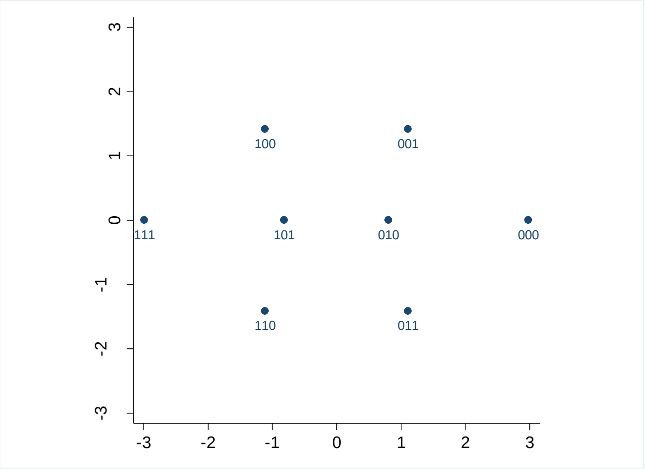
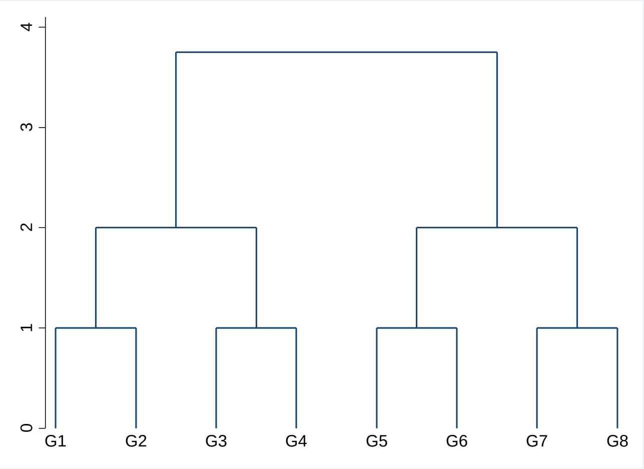


Fig.3. MDS *INDEL*=1.0; SUB=1.0; Fit(1)=0.56; Fit(2)=0.72

Fig.4. MDS *INDEL*=1.0; SUB=2.0; Fit(1)=0.70; Fit(2)=0.93



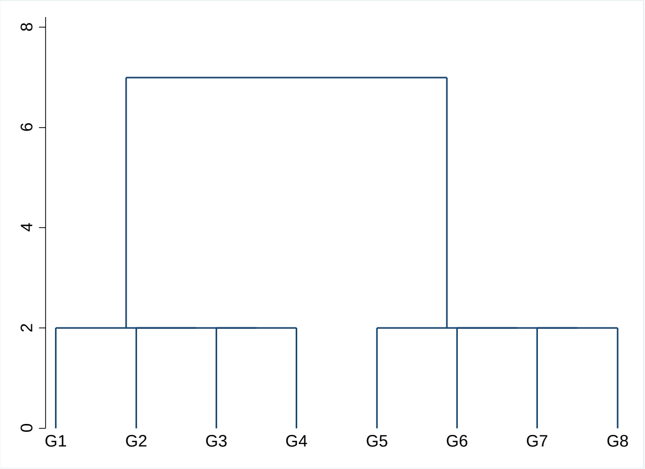


Fig.5. Hierarchical Cluster *Indel*=1.0; SUB=1.0

Fig.6. Hierarchical Cluster *INDEL*=1.0; SUB=2.0

I believe that Figures 1., 2., 3. and 4. do not require comment. For values less than 1.0, the analyses did not find an acceptable bi-dimensional solution of the eight sequences. The fit of the model was below the acceptability threshold. However, if we want to interpret the configuration of the points, we may observe that for a substitution cost of 0.5 the MDS arrange the sequences (Figure 1.) in a circle with the sequences {000 and 111} at its centre. For a value of 0.6, therefore, for a cost little higher than the previous one, the graph (Figure 2.) changes completely and draws two clusters of complementary sequences. The first, at upper right, consists of the sequences {011, 110, 000, 010} and the second, at bottom left, consists of the sequences {100, 001, 111, 101}.

For substitution costs greater than or equal to 1.0, acceptable solutions are found, but they are very different; with a cost of 1.0 (Figure 3.) the graph is almost the same as the graph obtained with a substitution cost of 0.5 (Figure 1.). The sequences are arranged in a circle with the complementary sequences {010, 101} perfectly overlapped at its centre. Nevertheless, the best solution (Figure 4.) is that for a substitution cost of 2.0.

It is therefore clear that varying the substitution costs produces radically different MDS models. Obviously, this may induce the researcher to draw very different conclusions according to the substitution cost adopted.
The situation does not improve when changing the type of analysis. The second analysis reported here was a classic hierarchical clusters analysis (Figures 5. and 6.). For brevity, I analyzed the distance matrix obtained for a substitution cost of 1.0 and 2.0 with a "Wards Linkage method". In this case, too, the solutions depicted in the dendogram of Figures 5. and 6. do not require comment. For a value of 1.0, four clusters of distinct sequences are found; while for a value of 2.0 there is no cluster of sequences.

Conclusion

This article has investigated the interaction effects exerted by substitution and *indel* costs on OM results. It has sought to answer a simple question: is it possible that the choice of the substitution and *indel* costs made by the researcher modifies the reality under examination? The conclusion is 'yes'. It has been shown that variation of substitution and insert/delete costs yields different solutions that lead to very different conclusions. The first group of analyses on the 'Ideal Type' showed that, all *indel* costs being equal, radically different solutions are obtained, both when (1) maintaining the reference sequence unchanged and varying the substitution costs, and (2) maintaining the substitution cost constant and making the reference sequence vary.
It has therefore not been possible to find any regularity,[4] or more simply any general rule, that guarantees the stability of the results. This raises several problems during the analysis phase, because is impossible to distinguish the real clusters occurring in reality from those that only result from distortions introduced into the analysis by the OM distances calculation. What is the consequence? It is that nothing tells us that the outcome obtained by analysis with 'ideal types', even approximately, is the 'real' configuration of the observed reality, and not the distortion effect introduced into the analysis by using one *indel* or substitution cost rather than another.

The situation does not improve with the standard OMA. In this case, too, the distance between sequences does not vary proportionally to the variation of costs. The result is that fixing the *indel* costs and varying the substitution costs yields very different distances matrixes not reducible to a common structure. Nor does the use of data reduction algorithms like Multidimensional Scaling and Hierarchical Cluster provide a solution. It is evident from the simulation that varying costs changes the solutions and radically alters the conclusions that the researcher can derive from the results.

Detailed analysis shows that, between the two statistical methods, the MDS is more sensitive to the variation of substitution costs. Whereas the Wards Linkage method finds the same clusters (for the eight sequences) independently of the value of the substitution costs, provided that costs are less than or equal to the *indel* cost, the results are totally different for a substitution cost double the *indel* cost. Nevertheless, we do not know what happens if hierarchical cluster methods other than the Wards method are used.

To conclude, the analyses reported here were very simple and based only on binary strings with a single substitution cost. But reality is much more complex: consider when several states and substitution costs are defined in the same analysis. What happens when, in the same analysis, the value of *indel* is fixed but different substitution costs are defined for the different states, or when the substitution cost is fixed and the *indel* varies, or when both costs vary? What solutions do we obtain?

[4] The only regularity found is with identical SI rates. This is, for instance, the solution obtained for a value of *indel*=1.0 and substitution=0.5 is the same as had with a cost of *indel*=10 and a substitution cost=5.0. In fact both these costs have exactly the same SI rate.

I believe that I have shown clearly the problems, risks and instability attendant on the use of OM techniques in sequence analysis. It is evident that only a simple small variation in substitution costs produces very different solutions. Moreover, these solutions are not comparable with each other; nor are they reducible through techniques of analysis to stable solutions.

What is the risk? It is the impossibility of clearly separating the observed reality from the artificial reality created by varying substitution and *indel* costs. What conclusions will a researcher draw if s/he obtains different solutions on varying costs? Which reality will return us these analyses? Will it be that we want to capture by sequences analysis or will it be that "artificial" obtained from the OM technique?

References

Aasave, A., Billari, F. and Piccarreta, R. (2007) 'Strings of adulthood: Analyzing work-family trajectories using sequence analysis', *European Journal of Population*, 23(3-4), 369–388.

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. (1984) 'Event sequence and event duration: Colligation and measurement', *Historical Methods*, 17(4), 192–204.

Abbott, A. (1988) 'Transcending general linear reality', *Sociological Theory*, 6, 169–186.

Abbott, A. (1990) 'Conceptions of time and events in social science methods', *Historical Methods*, 23(4),140–150.

Abbott, A. (1991a) 'History and sociology: the lost synthesis', *Social Science History*, 15(2), 201–238.

Abbott, A. (1991b) 'The order of professionalization: an empirical analysis', *Work and Occupations*, 18(4), 355–384.

Abbott, A. (1992) 'From causes to events: Notes on narrative positivism', *Sociological Methods and Research*, 20(4), 428–455.

Abbott, A. (1995) 'Sequence analysis: New methods for old ideas', *Annual Review of Sociology*, 21, 93–113.

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

Abbott, A. (2001), *Time Matters: On Theory and Method*. The University Chicago Press: Chicago.

Abbott, A. and DeViney, S. (1992) 'The welfare state as transnational event: Evidence from sequences of policy adoption', *Social Science History*, 16(2), 245–274.

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 analysis of musicians' careers', *American Journal of Sociology*, 96(1), 144–185.

Abbott, A. and Tsay, A. (2000) 'Sequence analysis and optional matching methods in sociology', *Sociological Methods and Research*, 29(1), 3–33.

Anyadike-Danes, M., and McVicar, D. (2003) *Parallel Lives: Birth, Childhood and Adolescent Influences on Career Paths*. Belfast: Northern Ireland Economic Research Centre. http://ideas.repec.org/p/ecm/ausm04/134.html

Anyadike-Danes, M. and McVicar, D. (2005) 'You'll never walk alone: Childhood influences and male career path clusters', *Labour Economics*, 12(4), 511–530.

Blair-Loy, M. (1999) 'Career patterns of executive women in finance: An optimal matching analysis', *American Journal of Sociology*, 104(5), 1346–1397.

Brzinsky-Fay, C., Kohler, U., & Luniak, M. (2006) 'Sequence analysis with Stata', *Stata Journal*, 6(4), 435-460.

Chan, T.W. (1995) 'Optimal Matching Analysis: A methodological note on studying career mobility', *Work and Occupations*, 22, 467–490.

Chan, T.W. (1999) 'Optimal Matching Analysis', Social Research Update University of Surrey, issue 24, http://sru.soc.surrey.ac.uk/SRU24.html.

Clark, W.A.V., Deurloo, M.C. and Dieleman, F. (2003) 'Housing careers in the United States, 1968-93:

Modelling the sequencing of housing states', *Urban Studies*, 40(1), 143–160.

Clote, P. and Straubhaar, J. (2006) 'Symmetric time warping, Boltzmann pair probabilities and functional genomics', *Journal of Mathematical Biology*, 53, 135–161.

Durkheim, É. (1912) Les Formes Élémentaires De La Vie Religieuse, F. Alcan, Paris, Italian translation: Le forme elementari della vita religiosa, Edizioni Comunità, Milano (1971).

Elzinga, C.H. (2003) 'Sequence similarity: A non-aligning technique', *Sociological Methods and Research*, 32(1), 3–29.

Elzinga, C.H. (2005) 'Combinatorial representations of token sequences', *Journal of Classification*, 22(1), 87–118.

Rhower, G. and Potter, U. (2009) *TDA User's Manual*, Ruhr-Universität, Bhocum http://steinhaus.stat.ruhr-uni-bochum.de/tman.html.

Halpin, B. and Chan, T.W. (1998) 'Class careers as sequences: An optimal matching analysis of work life histories', *European Sociological Review*, 14(2), 111 – 130.

Han, S.-K. and Moen, P. (1999) 'Work and family over time: A life course approach', *Annals of the American Academy of Political and Social Science*, 562, 98–110.

MacIndoe, H. and Abbott, A. (2004) Sequence Analysis and Optimal Matching Techniques for Social Science Data. In M. Hardy and A. Bryman (eds.) *Handbook of Data Analysis*. London: Sage Publications, 387–406..

Kruskal, J.B. and Liberman, M. (1983) The symmetric time-warping problem, in Sankoff, D. and Kruskal, J. B., 125–161.

Lesnard, L. (2006) *Optimal matching and social sciences, Document du travail du Centre de Recherche en Économie et Statistique 2006-01*. Institut Nationale de la Statistique et des Études Économiques: Paris.

Levenshtein, V.I. (1966)' Binary codes capable of correcting deletions, insertions, and reversals', *Soviet Physics Doklady*, 10, 707-710. Originally published in Russian in *Doklady Akademii Nauk SSSR*, 163 (4): 845-848, 1965.

Levine, J.H. (2000) 'But what have you done for us lately? Commentary on Abbott and Tsay', *Sociological Methods and Research*, 29(1), 34–40.

Levy, R., Gauthier, J.A. and Widmer, E. (2006) 'Entre contraintes institutionnelle et domestique: les parcours de vie masculins et feminins en Suisse', *Canadian Journal of Sociology*, 31(4), 461–489.

Malo, M.A. and Muñoz-Bullón, F. (2003) 'Employment status mobility from a life-cycle perspective: A sequence analysis of work-histories in the BHPS', *Demographic Research*, 9, 119–162.

Marteau, P.F. (2007) *Time Warp Edit Distance with Stiffness Adjustment for Time Series Matching*, ArXiv Computer Science e-prints.

Martin, P., Schoon, I. and Ross, A. (2008) 'Beyond Transitions: Applying Optimal Matching Analysis to Life Course Research', *International Journal of Social Research Methodology*, 11(3), 179-199.

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 (Series A)*, 165, 317–334.

Mead G.H., (1932) *The Philosophy of the Present*, Chicago: Open Court Pub. – Republished in 2002 by Prometheus Books: New York.

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: Series A*, 170(1), 167–183.

Pollock, G., Antcliff, V., & Ralphs, R. (2002) 'Work orders: Analysing employment histories using sequence data', *International Journal of Social Research Methodology*, 5, 91–105.

Reilly, C., Wang, C. and Rutherford, M. (2005) 'A rapid method for the comparison of cluster analyses', *Statistica Sinica*, 15(1), 19–33.

Scherer, S. (2001) 'Early career patterns: A comparison of Great Britain and West Germany', *European Sociological Review*, 17(2), 119–144.

Spencer H. (1855) *The Principles of Psychology*. London: Longman, Brown, Green and Longmans.

Sankoff, D., Kruskal, J.B. (eds.) (1983) *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*, Reading, MA.: Addison-Wesley.

Stovel, K. and Bolan, M. (2004) 'Residential trajectories: Using optimal alignment to reveal the structure of residential mobility', *Sociological Methods and Research*, 32(4), 559–598.

Stovel, K., Savage, M. and Bearman, P. (1996) 'Ascription into achievement', *American Journal of Sociology*, 102, 358–99.

Waterman, M.S. (1995) *Introduction to Computational Biology*, London: Chapman and Hall.

Wiggins, R.D., Erzberger, C., Hyde, M., Higgs, P., and Blane, D. (2007) 'Optimal matching analysis using

ideal types to describe the lifecourse: an illustration of how histories of work, partnerships and housing relate to quality of life in early old age', *International Journal of Social Research Methodology*, 10(4), 259-278.
Wilson, C. (2006) 'Reliability of sequence-alignment analysis of social processes: Monte Carlo tests of ClustalG software', *Environment and Planning A*, 38(1), 187.
Wu, L.L. (2000) 'Some comments on "Sequence analysis and optimal matching methods in sociology: Review and prospect"', *Sociological Methods and Research*, 29(1), 41–64.