University of Limerick
Department of Sociology Working Paper Series

Working Paper WP2008-01
January 2008

Brendan Halpin

Department of Sociology

University of Limerick

Optimal Matching Analysis and Life Course Data: the importance of duration

Available at `http://www.ul.ie/sociology/pubs/wp2008-1.pdf`

Optimal Matching Analysis and Life Course Data: the importance of duration

Brendan Halpin
Dept of Sociology, University of Limerick*

January 2008

Abstract

The Optimal Matching Algorithm is widely used for sequenace analysis in sociology. It has a natural interpretation for discrete-time sequences, but is also widely used for life history data, which is continuous in time. Life history data is arguably better dealt with in terms of episodes rather than as a string of time-unit observations, and the paper addresses the question of whether standard the OM algorithm is unsuitable for such sequences. A modified version of the algorithm is proposed, which weights OM's elementary operations inversely with episode length. In the general case, the modified algorithm produces pairwise distances much lower than the standard algorithm, the more the sequences are composed of long spells in the same state. However, where all the sequences in a data set consist of few long spells, and there is low variability in the number of spells, the modified algorithm generates an overall pattern of distances that is not very different from standard OM.

1 The sociological meaningfulness of the optimal matching algorithm

This paper addresses the Optimal Matching algorithm (OM), a very common approach to sequence analysis of life course trajectories, and explores the extent to which its operations can be considered to be sociologically meaningful. Particular attention is paid to the issue of using a technique designed for discrete sequences, for trajectories such as life courses or employment histories, which are better considered as taking place in continuous time. A modified algorithm, which introduces a sensitivity to spell length, is introduced and tested – it produces inter-sequence similarities which differ systematically from the standard algorithm, but makes surprisingly little difference to analysis of real data sets, suggesting that, though not designed for continuous-time sequences, standard OM is reasonably robust as long as the variability in spell length is moderate.

Following years of enthusiastic evangelism from Andrew Abbott and colleagues (Abbott, 1983, 1984, 1988, 1990; Abbott and Hrycak, 1990; Abbott, 1991a,b; Abbott and DeViney, 1992; Abbott, 1992, 1995; Abbott and Tsay, 2000; Abbott, 2000), the use of sequence analysis – in particular the optimal matching algorithm – is seeing a steady increase in application in the social sciences (*inter alia*,

*I am grateful for feedback at presentations in the Geary Institute, UCD, April 2007, the SAI Annual Conference, May 2007, and the WZ-Berlin Workshop on Sequence Analysis, September 2007. Contact: Brendan Halpin, Dept of Sociology, University of Limerick, `brendan.halpin@ul.ie`. {Version: omadjdur.tex,v 1.5 2008/01/31 22:49:30 brendan Exp}

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; Pollock, 2007; Aasave et al., 2007). However, the use of OM is not without controversy. It has been broadly criticised as having had little success in applications (e.g., Levine, 2000) and rather more acutely as being sociologically meaningless (Wu, 2000). The general question of successful application is in fact an empirical one, and as more papers are published using the method we see a real advantage in the ability to access sequence information holistically, if not to the degree of overthrowing "general linear reality" (Abbott, 1988). That is to say, there are many genuinely effective but not paradigm-shifting applications of OM in the sociological literature. However, the issue of the sociological meaningfulness of the technique remains of great importance – we need to have a clear understanding of how (and indeed whether) it extracts sociological information from the data; it may be that successful applications work despite the technique rather than because of it, but more optimistically we can say that a precise understanding of how it works is necessary to apply it more successfully.

Wu articulated a sustained critique of the OM algorithm in debate with Abbott (Abbott and Tsay, 2000; Wu, 2000; Abbott, 2000). He noted the success of alignment techniques in molecular biology contexts, and attributed this at least in part to the close match between the OM algorithm's "elementary operation" of substitution with the biological process of mutation at a site in a DNA sequence. He proceeded by arguing that the application of the same techiques to temporal sociological sequences was hampered, if not invalidated, by the much weaker analogy between substitution as an operation on a sequence and the temporal processes of change involved in the creation of life-course trajectories. I fully endorse his requirement that the method be sociologically meaningful, but I feel his analysis was marked by a crucial misunderstanding of the meaning of the elementary operations of the algorithm. In particular, his insistence on regarding substitutions as representing transitions, which underlies a large part of his critique, is completely inappropriate. While transitions are events that occur in time, within a sequence, substitutions are atemporal operations that involve comparing two sequences at a particular site. Many of the problems he raises, such as that of "impossible" transitions or substitutions, or the asymmetry between comparing a transition to unemployment with a transition from unemployment, are artefacts of this misunderstanding. All the substitution operation implies is that two sequences are (in part) dissimilar to the extent that the states in the pair are dissimilar. Temporality or sequence linearity only comes in via the repeated execution of the operation on the sequences – the sequences contain all the linear or temporal information, not the individual elementary operations.

While Wu was incorrect to identify substitutions as transitions, his concern with transitions as substantively important is well founded. Techniques which focus on transitions, particularly the family of hazard-rate modelling approaches, are in many respects superior to holistic sequence analysis. They allow us to model longitudinal processes in terms of their generative logic, and are therefore more capable of answering specific theoretical questions, whereas sequence analysis as it currently exists tends to be more descriptive and exploratory in utility. Sequence analysis is better seen as complementing more conventional strategies than competing with them. In particular, it allows the researcher to apprehend the overall structure of complicated longitudinal data, and it gives a holistic perspective that can help put the spell-focused hazard rate model, or the period-by-period transition-focused model, into context.

2 The logic of comparison

We start from a sociologically derived understanding of differences within a state space. We wish to move to a sociological assessment of differences between trajectories (which we can assume are on a meaningfully comparable time axis, e.g., representing comparable parts of life courses). The OM strategy to achieve this has three main elements: substitution, which says that the distance between two sequences which differ in a single location is related to the difference between the differing states; alignment, such that sequences that match in different locations are dissimilar to the extent that the locations are apart; and a minimising cumulation, such that the difference between any two sequences can be calculated as the "cheapest" concatenation of substitution and alignment operations which maps one to the other.

Taking the three points in reverse order, the minimising concatenation of operations seems an intuitively satisfactory means of translating the elementary operations into a distance between any pair of sequences. It of course depends on the meaningfulness of the elementary operations, as well as on additivity: if we can change s_1 into s_2 by deleting, say, $s_{1,i}$ to create s'_1, and then using substitution to change $s'_{1,j}$ from α to β (and there is no cheaper route), we calculate to total distance thus:

$$D(s_1,s_2) = D(s_1,s'_1) + D(s'_1,s_2) = \delta + \sigma_{\alpha,\beta}$$

While it seems intuitively uncontroversial that the optimum total cost should be derived from the "cheapest" route, it is quite possible to conceive of different ways of cumulating the costs across the intermediate comparisons. For instance, $D(s_1,s'_1) + D(s'_1,s_2)$ could stand as a maximum for $D(s_1,s_2)$ rather than as an exact calculation, if we were to conceive of s_1, s'_1 and s_2 as constituting a triangle. However, the additive form has the great virtue of conceptual simplicity and ease of implementation.

Alignment is similarly intuitively attractive, such that the distance between two sequences with a partial match is related to how far apart the match is. We can see limitations in the procedure, in that alignment will target only a single longest subsequence. If a pair of sequences have two submatches that are in alternate order, for instance, `xABCxDEFx` and `yDEFyABCy`, alignment is not capable of accounting for the dual match.[1] Nonetheless, the ability to recognise similarity when it occurs "out of phase" is particularly useful.[2]

Substitution is a very important aspect of the OM approach, an adaptability that increases its sociological utility. Here by substitution I mean the practice of costing certain substitutions as less than the "ceiling" cost of one deletion followed by one insertion, that is, the use of a matrix of pairwise substitution costs. We do not just count and align perfect matches, but we also have a means of assessing mismatches differentially, on the basis of our knowledge of the original, non-sequential, state space. While it is often reasonable to propose a state space structure such that all state-pairs are equally dissimilar, we will often be able to describe state spaces – be it informed by intuition, theory or data – such that some pairs of states are more similar than other pairs (for instance, we may judge training and education to be more similar to each other than either is to unemployment). That is,

[1] "All common subsequence" algorithms, for instance those of Elzinga (2005, 2003) will register this extra similarity (and furthermore correctly identify the even greater similarity of `xABCxDEFx` and `yABCyDEFy`), but are outside the scope of the current paper.

[2] Sometimes, of course, being out of phase is a substantively important dimension of difference (see for instance, Lesnard, 2006). In such cases, *indel* operations must be made more costly or suppressed altogher (resulting in a Hamming distance comparison).

while *indel*s will cost the difference between ABC and ADC as 2δ, the use of pair-specific substitution costs allows us to rank, for instance ABC and ADC as more similar than ABC and AEC, on the basis that we have previously determined B to be more similar to D than to E.[3]

As can be seen, substitutions do not in any way imply transitions (which are longitudinal in orientation) but rather a lateral comparison between elements in separate sequences. The essential quality here is similarity (or difference) between the pair of states, which is unproblematically symmetric, and does not logically involve concepts of transition or arrows of time, *pace* Wu (2000). To say, for instance, that the state of "never-married" is closer to "divorced" than to "married" (if single–partnered is an important dimension in the analysis) has no implication whatsoever that divorced people should become never-married.

It should be noted in passing that, quite independently of sequence analysis, transitions *may* provide a good way of defining inter-state differences, insofar as it is reasonable to use frequency of movement between states as an inverse measure of distance. However, for some sorts of state spaces, and some theoretical concerns, this may not be appropriate. If a state space can be considered as a partitioning of a latent multi-dimensional space, where the probability of a move from one location to another is inversely proportional to the distance between them, then transition probabilities do tell something useful about the structure of the partitioning. However, in many applications this sort of unstructured movement within the latent space may not be a good assumption. For instance, we can think of the relatively high rate of transitions between unemployment and employment as informing us more about the difference between them (in particular, the fact that unemployment is much less desirable than employment) than about the similarity. Other strategies for determining inter-state distances could include locating them in a specified multi-dimensional space, for instance by using other data sources to characterise the states in terms of a number of factors. For example, an occupational classification could be supplemented by information on, say, the average level of education and income and the level of gender segregation within each category, and inter-state differences calculated as a function of distances in the implied three-dimensional space. (Clearly, the sorts of factors adduced like this depends on the substantive focus of the analysis.) Another point to bear in mind is that some state spaces may include distinctions that are not immediately relevant, and affect the meaning of transition rates. For instance, if an occupational classification makes a distinction between male and female nurses, transition rates will suggest that these two very similar groups are maximally different.[4] In such a case, even if the later analysis requires the retention of the gender distinction, it may be appropriate to temporarily suppress it in calculating transition-based similarities.

Let us return to the plausibility of OM and its elementary operations. Starting from a sociologically meaningful understanding of differences within a state space, OM allows us to proceed to an understanding of differences between trajectories, via alignment, substitution and minimising cumulation, three elements which can be agreed to have at least *prima facie* plausibility. Substitution in particular, in defining the distance between ABC and ADC as a function of the B–D difference in the

[3] This is an important advantage over common-subsequence methods as hitherto proposed (Elzinga, 2005, 2003), but such methods can readily be extended to accommodate variable substitution costs (Elzinga, personal communication)

[4] To return to the "impossible" transition is between never married and divorced: these states are not necessarily maximally dissimilar but should have no observed transitions. Here, a re-definition may help, such that sequential information is dropped from the state space, and allowed to reside entirely in the sequence. Thus we could define both states as "single/not-legally-married", allowing the distinction to reside in the fact that the divorced state comes after a legally-married spell. All the information we lose in this re-definition is retained in the sequence (assuming it is long enough).

initial state space, is clear and simple. However, it involves assumptions that may not always be acceptable. For instance, the distance between ABC and ADC is necessarily the same as that between EBF and EDF, that is, the substitution cost is blind to the environments of the pair of elements being considered. Similarly, substitution costs are the same whether the comparison is made early in the sequence or late (ABCCCCC and ADCCCCC are identically distant as AAAAABC and AAAAADC).[5] However, it can be argued that where the data show patterns such that, for instance both Bs and Ds are disproportionately likely to be found in the middle of A.C sandwiches (and thus in a particular sense, to be more like each other when found in particular environments), this will properly emerge in the results of the analysis as an outcome (for instance, with Bs and Ds disproportionately found in clusters characterised by A.C). That is, the fact that there is a temporal or linear logic (that certain states are disproportionately likely to follow or precede other specific states) is a feature of the longitudinal nature of the trajectory rather than of the state space. In that much, it is better to emerge as a result of the sequence analysis than to be fed into it as an picture of the initial state space that already includes a strong longitudinal perspective.

There is one other important consideration here: OM is structured in terms of discrete-time sequences, as an ordered set of discrete elements, usually representing a single time unit. Many sociological applications will be correspondingly discrete, where the elements may be utterances in a conversation, brands bought in successive purchases, steps in a dance, and so on. However, a lot of sociological sequence analysis is carried out on processes that would much better be represented in continuous time, such as life courses or employment histories. The usual practice is to represent such sequences with spell lengths rounded to a relatively small time unit such as the month, with elements repeated proportionally to the length of the spell. In practice this works well, but in the next section I want to consider whether the discrete logic is innapropriate for such trajectories.

2.1 Shortcomings of OM: persistence in state

In contrast to naturally discrete longitudinal data, like sequences of utterances in a conversation, life course data is more conveniently represented in spell format, as a sequence of episodes of given duration and state. Purely discrete data must be represented simply as sequences of elements (which may have constant duration but may be simply logically separate), and this is the format which OM sequence analysis uses. Spell data can be structured for OM in a number of ways. The simplest is to represent the spells as elements, ignoring duration. However, differences in duration are likely to be substantively significant so this is usually unsatisfactory, and the more common approach is to represent spells as strings with one element per time unit (for example, Abbott and Hrycak, 1990; Halpin and Chan, 1998; Blair-Loy, 1999; Scherer, 2001; Malo and Muñoz-Bullón, 2003; Anyadike-Danes and McVicar, 2005; Pollock, 2007). As the literature attests, this strategy yields useful results. However, this is not a natural way to treat spell data.[6] With discrete data, operations directly on individual elements make sense, but with discretely-represented continuous-time data indels and substitutions instead affect artificially defined sub-portions of spells. This raises problems with the sociological meaningfulness of the elementary operations. In the discrete case, substitution is meaningful insofar

[5](Lesnard, 2006) reports the application of "dynamic Hamming distances" in time diary analysis, where he makes a strong case that substitution costs should vary.

[6]Elzinga's combinatorial methods (2003; 2005) include approaches that deal with spells as elements, weighted by duration, which is attractive. However, the logic of these methods is sufficiently distinct from OM to be outside the scope of the present paper.

as the ABC–ADC distance is a function of the B–D distance, and is independent of the A.C environment. When we have spells represented as more-or-less long runs in the same state, this assumption of independence of the adjacent states is much less acceptable, because of the high probability of an element following or preceding an element in the same state. Thus OM tells us that s_1 = AAAB is as distant from s_2 = AACB as from s_3 = AABB (given $\sigma_{A,B} = \sigma_{A,C}$), while sociologically it is clear that s_1 and s_3 are distinctly more similar to each other than to s_2. Sociologically s_1 can be changed into s_3 by a relatively small change in timing, whereas s_2 contains a completely new spell in a different state. Similarly, from a sociological point of view, the deletion of a month from an 18-month spell is far less consequential than from a 2-month spell, but OM cannot recognise this. Depending on the application, this may be a serious problem. This is a case of "environment dependence" that is logically distinct from the discrete ABC–ADC example above, because the "environment" is not specific states such as A.C but rather the same state repeated, that is, persistence in state.

In the next section I propose a modified form of the OM algorithm, that efficiently takes spell length into account when costing the elementary operations.

3 Amending OM to be sensitive to spell length

We can characterise the problem most simply in terms of deletion (insertion is equivalent to deletion in the other sequence, and substitution involves a deletion and an insertion, so what follows does not lose generality). In the discrete case, if our elementary operation requires a deletion, the environment of the deleted element is not important. In the pseudo-continuous case (i.e., representing spell data as runs of time-unit elements) the environment can be seen to matter, in that sociologically the importance of shortening a spell by one unit will vary inversely with the length of the spell. That is, we might wish the cost of deleting a unit to be less in a long spell than in a short spell.

More generally, the cost of deleting a single element in a spell should:

- be the same as standard OM if the spell is one unit long,
- be lower the longer the spell,
- but sufficiently high that deleting all of a longer spell costs more than all of a shorter spell.

One way of achieving this could be to generate the pseudo-discrete spells in a non-linear fashion, perhaps using the log of the length of the spell to calculate the number of elements required to represent it (more specifically, $round(log(l+1))$ where l is spell length, in order to generate non-zero pseudo-lengths for short spells). In this fashion, removing an element from a shorter spell represents a smaller amount of real time (but still a larger fraction of the spell). However, in the case of logs, there are problems of discretisation, insofar as spells of quite different true length will be represented with the same pseudo-length (for instance, spells with true lengths between 4 and 11 have a pseudo-length of 2). There are other functions, however, which do not suffer too badly from discretisation, and much of what follows could be approximated by restructuring the data with a modified spell length defined as $round(\sqrt{l \times L})$ where l is spell length and L is a number a little larger than the maximum spell length.

Substitution costs

s_2 \ s_1	A	B	C	D
C	2	1	0	1
D	3	2	1	0
A	0	1	2	3
A	0	1	2	3
B	1	0	1	2

OM workspace

s_2 \ s_1					
	0	2	4	6	8
2	2	2	3	4	6
4	4	4	4	4	4
6	4	5	6	6	6
8	6	5	7	8	8
10	8	6	6	8	8

Figure 1: The OM calculations and workspace

3.1 The OM algorithm in detail

The strategy I present here adapts the OM algorithm to adjust substitution and indel costs according to spell length. This can be done with relatively little impact on the efficiency of the algorithm. I begin by outlining the standard operation of the OM algorithm, to put the changes in context. OM uses dynamic programming techniques to calculate the cheapest set of elementary operations to transform one sequence into another – it does this in a maximally efficient manner, hence "optimal". Its operation can be represented as taking place in a workspace in the form of matrix of dimension $(l+1) \times (m+1)$ where l and m are the sequence lengths. The top left cell is given the value 0, and the rest of the first row and column are filled with multiples of the indel cost. The remaining cells are then filled in an iterative procedure, according to the formula:

$$C_{ij} = min \begin{cases} C_{i-1,j-1} + \sigma_{i,j} \\ C_{i,j-1} + \delta \\ C_{i-1,j} + \delta \end{cases}$$

where δ is the indel cost, and $\sigma_{i,j}$ is the substitution cost involved in swapping between the ith element of sequence 1 and the jth element of sequence 2 (see Figure 1). That is, the cell value is the minimum of three possibilities:

- the cell above-left, plus the substitution cost,
- the cell above plus the indel cost, and
- the cell to the left plus the indel cost.

At this point, the algorithm is looking at a specific pair of locations and calculating whether substitution, insertion or deletion is the cheapest operation. To give a concrete example, consider the sequences, s_1 = `ABCD` and s_2 = `CDAAB`. For simplicity, I define substitution cost as the distance between the letters (i.e., `A` to `D` is 3), and let the indel cost be 2 (such that two indels cost slightly more than the highest substitution cost).

Rows 2 to $l+1$ and columns 2 to $m+1$ represent the comparison of all possible pairs of the two sequences. Cell (2, 2) thus represents the comparison of `A` in s_1 with `C` in s_2. We can resolve the inconsistence by substitution, at a cost of $C_{1,1} + \sigma_{A,C} = 0 + 2 = 2$ or by an insertion–deletion pair, involving first moving to $C_{1,2}$ (or equivalently $C_{2,1}$) and then to $C_{2,2}$, at a cost of $2 + 2 = 4$. Substitution gives the minimum, hence $C_{2,2}$ is assigned the value 2. $C_{2,3}$ is then calculated, and the cheapest way of getting there is to do a `B-C` substition from $C_{1,2}$. The process continues in that manner,

Substitution costs

s_2 \ s_1	A	B	C	D
C	2	1	0	1
D	3	2	1	0
A	0	0.7	1.4	2.1
A	0	0.7	1.4	2.1
B	1	0	1	2

OM workspace

s_2 \ s_1	0	2	4	6	8
2.0	2.0	3.0	4.0	6.0	
4.0	4.0	4.0	4.0	4.0	
5.4	4.0	4.7	5.4	5.4	
6.8	5.4	4.7	6.1	6.8	
8.8	7.4	5.4	5.7	7.7	

Figure 2: The modified OM calculations and workspace

with subsequent comparisons taking account of what went before. Thus the comparison in cell (2, 3) finds that the cheapest option is the $c_{1,2} + \sigma_{B,C}$, implying the cheapest partial solution so far involves deleting the A and turning the B into a C. By the time the algorithm reaches the bottom right cell, all possible comparisons have been implicitly considered, and the value in that cell constitutes the cost of the cheapest route from s_1 to s_2 (this is often standardised by dividing by the length of the longer sequence).

3.2 Modifying OM

Modifying the procedure to cost operations on spells differentially according to their length is straightforward. First, we decide on the function by which we discount spell length. Many functions will satisfy the conditions mentioned above, but a simple strategy is to divide the cost by the square root of the length of the spell. The cost of deleting an element thus falls with length (i.e., $\frac{1}{\sqrt{x}}$ falls with x) but the cost of deleting a whole spell rises with length ($\frac{x}{\sqrt{x}}$ rises with x). Further, there is no effect on a 1-unit spell.

There are three sorts of operation, but the justification for this procedure has been given only in terms of deletion. This is not a problem as, first, insertions and deletions are interchangeable. The same result is achieved by inserting an element in one sequence as by deleting the corresponding element in the other sequence. Second, substitutions involve a deletion and an insertion (though at a reduced cost). We can implement the duration-sensitive version by calculating the cumulated substitution costs (in the first row and column of the workspace) taking account of runs of the same value, and by adapting the substitution cost similarly (in what follows, the substitution cost is reduced according to the longer of the two spells it affects; using the arithmetic or geometric mean of $\frac{1}{\sqrt{r_1}}$ and $\frac{1}{\sqrt{r_2}}$, where r is spell-length, will have very similar effects, particularly where spell lengths are broadly similar).

Figure 2 shows the calculations for s_1 = `ABCD` and s_2 = `CDAAB` under the modified algorithm. Sequence 2 contains one repetition: consequently the cumulative indel costs in column 1 of the workspace differ, rising by 1.4 ($= 2 \times \frac{1}{\sqrt{2}}$) instead of 2. Similarly, the substitution costs affecting those elements are reduced by the same factor. The final result of the calculation is now 7.7 rather than 8.0, because of the presence of the two-element spell. If there had been no repetition, the result would have been 8.0 as before. That is, the modified OM algorithm (hereafter, OMv) generates distances less than or equal to the standard algorithm. Table 1 gives more examples, demonstrating that distances between sequences tend to fall, the more they consist of repeated elements. The differences are marked: while the first pair of sequences have no repetition and are given the same distance by

Table 1: OM and OMv distances for example sequences

Sequences		Distances	
A	B	OM	OMv
ABCDC	BCDAD	1.00	1.00
ABCDC	BBBAC	1.00	0.83
ABCDC	BBBBB	1.00	0.45
BCDAD	BBBAC	0.80	0.55
BCDAD	BBBBB	1.20	0.54
BBBAC	BBBBB	0.40	0.18

Note: substitution and deletion costs as in Figure 1; final costs are divided by sequence length.

both algorithms, pairs 4 and 5, which feature high levels of repetition, have OMv distances less than half the OM distance. This is entirely as desired, in that the longer the spell, the lower the cost of changing its length, and therefore, *prima facie*, the results are substantively more appropriate. However, the question still remains to what extent will the adapted measure perform differently with real data. Section 4 explores this question, first with simulated data that replicates aspects of the structure of life course data better than these short examples, and then with real life course data.

3.3 Varying the level of adjustment

The adjustment presented above is by a factor of $\frac{1}{\sqrt{r}}$, or $r^{-0.5}$ where r is the spell or run length. Other functions will fulfill the criteria outlined on page 3 equally well, but it is worth considering here one set, of which OMv as presented is a specific case. That is, the cost adjustment factor can be defined as $r^{-\lambda}$, with $\lambda = 0.5$ giving us the present version. For $0 < \lambda < 1$ the criteria are met; for $\lambda = 0$, $r^{-\lambda} = 1$ so there is no adjustment; for $\lambda = 1$, the cost of deleting any complete spell is the same, regardless of length; and for $0.5 < \lambda < 1$ the adjustment is stronger than in the $\frac{1}{\sqrt{r}}$ case.

4 Comparing OMv to OM

While the proof of the pudding is in the performance of the algorithm in real analysis of real data, a lot can be learnt by looking at its performance on simulated data, and on intermediate aspects such as the correlation between OM and OMv distances. I therefore begin by looking at how OM and OMv distances compare in simulated and real data sets, and then move on to carrying out a cluster analysis of 5-year employment histories of new mothers, comparing the empirical typologies derived by the two methods.

4.1 Correlation between OM and OMv

Table 1 suggests that the correlation between OM and OMv will be positive, but substantially less than 1. However, in a data set where all the sequences are characterised by long spells, and thus all the OMv distances will be much lower than OM, will it be the case that the overall pattern will not change much? To this end I present results on a set of simulated data sets with different distributions of spell lengths, and on two real data sets. The simulated data sets use a 4-category state variable, run

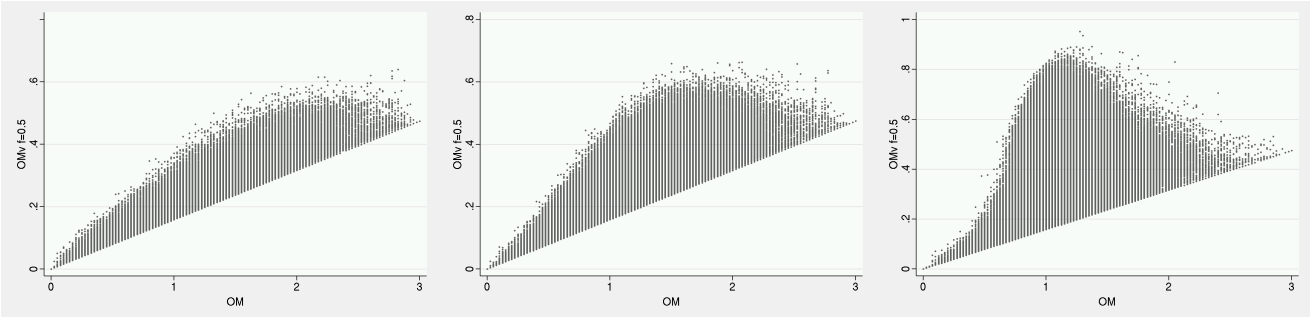

Figure 3: OMv and OM distances for simulated data: Low, medium and high numbers of spells

for 40 time units, and have a flat transition structure such that if a transition occurs, all destinations are equally likely. They differ in the base probability of transition, and thus in average spell length. The first simulated data set has about two spells on average, the second about 3.5 and the third 11. By virtue of the undifferentiated transition patterns they are quite unlike real life course data, but from the point of view of comparing OM and OMv they allow us to focus on the effect of average spell length. A simple substitution cost structure is used, where the distance between the values gives the substitution cost, and the indel cost is 2.

For the low-transition data set, the OMv distances between all pairs of sequence are on average only about one fifth of the OM distances, but the correlation is very high, at 0.963. All the sequences have relatively few, long, spells, so all OMv distances are reduced, but without changing the overall pattern radically. The medium data set, with about 3.5 spells per sequence rather than two, shows a similar drop on average distance, but nonetheless a high correlation at 0.818. The high transition data set, with about 11 spells per sequence, also shows a significant drop in the average distance (OMv distances at about 30% of OM) but the correlation is much smaller at 0.231. Sequences with shorter spells are affected by the "discounting" of elementary operation costs to a lesser degree, but they also have higher entropy (because of the more frequent transitions), and thus the choice of algorithm seems to matter more. By contrast, sequences with few spells are being discounted to a greater degree, but they are being compared against other low-entropy sequences and the choice of algorithm matters less. Figure 3 presents the relationships as scatterplots – all three panels suggest that there is a floor to the OMv distance of about 16% of the OM distance, but there is a good deal of variability above this floor. The high transition rate simulation has the most distinct pattern, with pairs whose OM value is in the middle of the range showing the most variation in their OMv distances.

Variability in the number of spells is also likely to be important – if all sequences tend to have similar numbers of spells of more or less similar length, pair comparisons will tend to be more often of like with like, than if the data consists of sequences of very different numbers of spells. The high-transition simulation has a particularly high variability, with 50% of cases having 8 or fewer spells, and 10% having more than 25 (the standard deviation of the number of spells is respectively 1.24, 2.57 and 9.03 in the three simulations, see Table 2).

The simulations give an important baseline, but it is necessary to see what happens with real data. Real data will be more complex than the simulations in a number of respects, perhaps most important being the transition patterns, which will be more complicated, may change over time, may be conditional on sequence history and observed and unobserved characteristics of the individuals. Two data sets are presented here, one consisting of monthly data on six years of women's labour

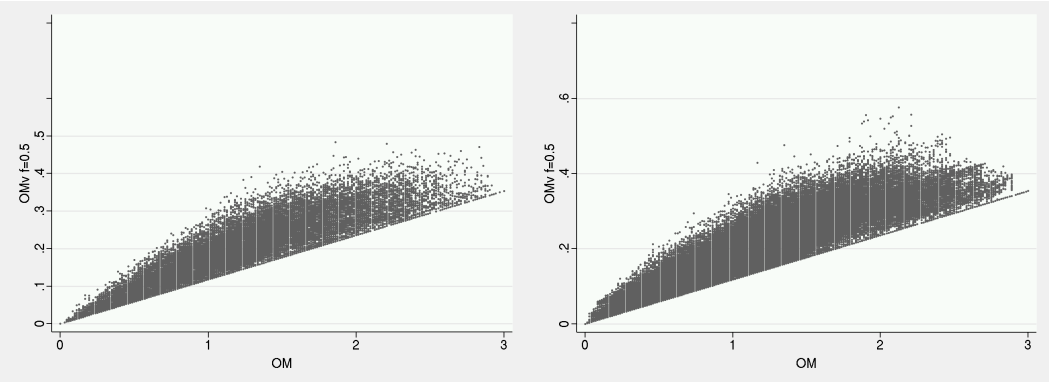

Figure 4: OMv and OM distances for mothers' labour market sequences (BHPS) and labour market entrants' sequences (MVAD)

market experience, drawn from the British Household Panel Survey, centred on the birth of a child in month 25, and the school-to-work transition data of McVicar and Anyadike-Danes (2002), also six years of monthly data. The BHPS data consists of four states (full-time employed, part-time employed, unemployed and non-employed) and a simple distance substitution matrix is used; the MVAD data has six states (school, further education, higher education, employment, joblessness) and their original substitution cost matrix is used. These are somewhat longer sequences than the simulations, with the same or similar numbers of states. The BHPS data has a very high correlation between OM and OMv, at 0.982, and the MVAD data a little lower at 0.924. Table 2 summarises the correlations, with additional information on the data sets, and with OMv at different λ settings as described in section 3.3 (the lower λ is, the closer OMv is to OM).

Both the real data sets are close to the low-transition data set in their level of correlation, and in the level of variability of number of spells, though the MVAD average number of spells is closest to the medium simulation. Nonetheless, it is clear that compared with truly discrete sequences of similar length, these sequences have quite low entropy, and as a result it seems that the modified distance measure does not produce dramatically different results. Figure 4 confirms that the pattern of the relationship between the measures is for both data sets quite similar to the low-transition simulation.

4.2 Varying λ

Table 2 also presents correlations for OMv with varied λ parameters. A value of 0.5 yields the standard algorithm as presented, values closer to zero yield versions with lesser discounting of long spells, and values closer to one greater discounting. For all five data sets, the correlations are monotonic in λ (the smaller the adjustment, the closer the scores to standard OM), but it is only for the high-transition simulation that we see radical disagreement, with the correlations close to zero for higher values of λ. For the two real data sets, correlations remain high even for the highest values. That is, for data with relatively few, longer spells, even a radical discounting of costs in the modified algorithm does not bring about a correspondingly radical change in the pairwise distance structure.

Table 2: Correlation of OM and OMv distances, using real and simulated data

		Correlation with OM distance				
		BHPS	MVAD	Simulations		
				Low	Med	High
OMv, $\lambda =$	0.1	1.000	0.997	0.999	0.994	0.953
	0.2	0.998	0.988	0.995	0.973	0.796
	0.5	0.982	0.924	0.963	0.818	0.231
	0.7	0.957	0.848	0.919	0.663	0.045
	0.9	0.913	0.742	0.851	0.508	-0.052
	1.0	0.881	0.676	0.808	0.438	-0.083
N cases		675	712	1000	1000	1000
Length		73	72	40	40	40
N states		4	6	4	4	4
Mean n spells		1.89	3.55	1.96	3.44	10.93
St.dev n spells		1.51	1.68	1.24	2.57	9.03

4.3 Cluster analysis of mothers' labour histories

While pairwise distances are the direct product of sequence analysis, the work does not usually stop there. Most typically the pairwise distance matrices are used to generate empirical typologies, data-driven classifications of the sequences. To explore how much modifying OM affects the outcome I present a cluster analysis of the BHPS maternal labour history data. To recap, these are six-year monthly labour market histories of women, who have a birth at the end of the second year, classified into full-time and part-time employment, unemployment and non-employment. The substitution costs imply a unidimensional linear structure to the state space (FT to PT is one, FT to UE is 2, FT to Non-E is 3, and so on), and the indel cost is 2. Table 3 presents the results of cluster analyses, using Ward's method and stopping at eight clusters, for OMv and OM pairwise distances. An eight-cluster solution is chosen on the informal grounds that it represents a manageable number of distinct clusters.

The results are so similar that it is quite easy to identify clusters across the two solutions, on the basis of maximal shared membership, but formally the mapping was generated by choosing the permutation of the OM cluster solution that maximised the κ score.[7] Remembering that the OM and OMv scores correlate at 0.982 it is not surprising that the cluster solutions should be close. 85.5% of cases are on the main diagonal and κ_{max} is 0.81. Indeed, one might even be surprised that with a correlation so high the match is not even closer. The more significant deviations from agreement are highlighted in red. The disagreement between the two measures has two major elements: OM splits nearly 40 cases out of OMv's largest cluster, mainly into two other clusters, and OMv spreads OM's third cluster across four clusters. Both sets of clusters are shown as index plots in Figure 5.

Not surprisingly, the biggest clusters consist of very high proportions of very simple sequences, with the simplest possible being six years in the same state. The biggest cluster consists mainly of mothers non-employed for the whole period, and permanent full-time employment characterises the next biggest. Part-time and unemployment clusters exist too but they are smaller. Broadly speaking,

[7] κ is a measure of the excess of cases on the diagonal over those expected under independence. Reilly et al. (2005) propose κ_{max} as a measure of agreement across cluster solutions, where κ_{max} is the highest κ across all permutations.

Table 3: OM and OMv 8-cluster solutions

OMv	OM								Total
	1	2	3	4	5	6	7	8	
1	263	28	0	0	0	10	1	0	302
2	0	39	7	0	2	0	0	0	48
3	0	0	18	0	0	0	0	0	18
4	0	0	19	54	1	0	0	0	74
5	0	0	0	0	33	0	0	0	33
6	0	0	0	0	0	13	0	0	13
7	0	0	0	0	0	3	18	0	21
8	0	0	27	0	0	0	0	139	166
Total	263	67	71	54	36	26	19	139	675

Note: Cells indicating substantial disagreement are highlighted in red.

that leaves four clusters characterised by change. However, it is notable that OMv is less reluctant than conventional OM to append non-static trajectories to the large no-change clusters. In both cluster solutions, cluster 2 is characterised by transitions from full-time to non-employment, more or less around the time of the birth. OM Cluster 3 is a mixture: exits from full-time to part-time around the time of the birth, later exits from full-time. Cluster 5 shows exits from full-time around or before the birth, a period of non-employment and then a return to part-time. Cluster 6 is small, with non-employment early and part-time employment later. As is often the experience with cluster analysis of sequence data, we get some large simple clusters of simple sequences, and some small complex clusters of more complex sequences.

But to focus on how OM and OMv differ, let us refer again to Table 3. The differences are broadly symmetric, in that in one case OMv splits up an OM cluster, and in another OMv makes a single cluster out of elements drawn from several OM clusters. To take the latter first, OMv cluster 1 draws 28 cases from OM cluster 2 and 10 from OM cluster 6. The agreed portion of cluster 1 consists almost exclusively of months in non-employment but OMv adds to that sequences characterised by early months in full-time or unemployment prior to consistent non-employment, and sequences characterised by a late transition from non-employment to part-time (see Figure 6). Clearly, the duration-sensitive algorithm rates these sequences as more similar to the no-transition sequences than OM does. This might be surprising, in that one might have expected they would have been judged as even closer to trajectories with the same sequences of states, but different durations.

The opposite happens with OM cluster 3, which is characterised by early full-time, but a complex mix of transitions and states later. OMv takes this small but relatively heterogeneous cluster and splits it in four. As Figure 7 shows, the subclusters are clearly more homogeneous than the full cluster, but we could have achieved this by moving further down the dendrogram. Where the subclusters go is more interesting. The first part, characterised by stepping down from full to part time around the birth, but then dropping out of the labour market, is moved to a cluster with a similar trajectory but without the intermediate part-time. The second block stays alone, and features labour market exit only after about four years. The third block is dominated by part-time (but far from 100% of the time) and is shunted into the 100% part-time block. The last block is very mixed but is dominated by full-time work with multiple transitions. This is also attached to the 100% full-time block.

It is difficult to draw conclusions from a single cluster comparison, to say which method is "sociologically" more adequate, particularly without first addressing the difficult question of what would

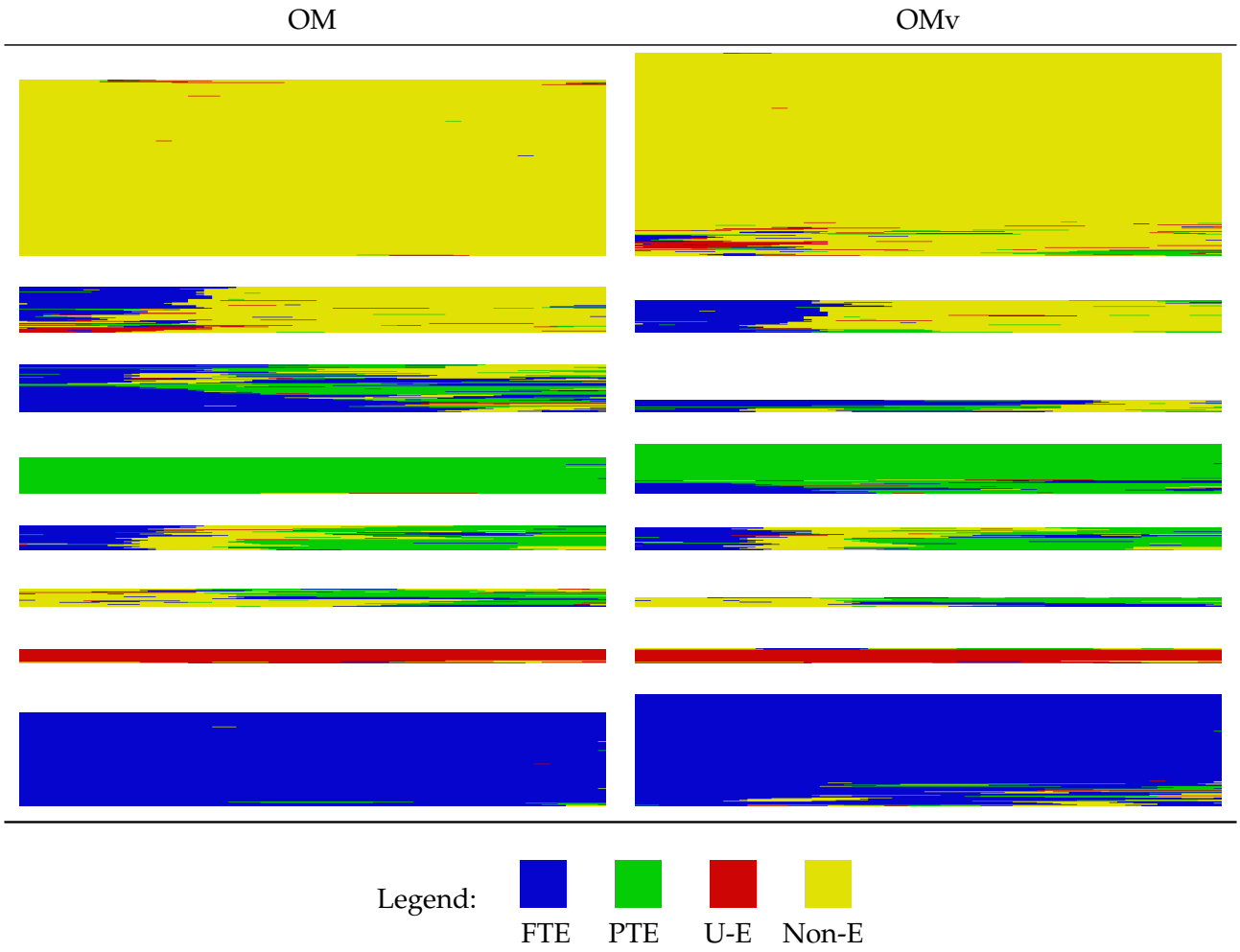

Figure 5: OM and OMv 8-cluster solutions compared, mothers' labour market data. Clusters are matched across the two solutions on the basis of maximal common membership. Each fine horizontal line represents one individual's history from month 0 to month 72.

constitute a "better" result. However, we can remark that OMv systematically finds single-state trajectories to be more like trajectories incompletely dominated by a single state, than conventional OM does. In doing this, it tends to be less likely to amalgamate these sequences in a rag-tag cluster. That said, perhaps the most important fact to take from the test is the relatively small effect on the final outcome.

5 Discussion

The duration-sensitive algorithm clearly produces different results, with far lower costs for pairs where one or both sequence has long runs of the same value. For data sets with high variability in number of spells per sequence, this produces a very different set of pairwise distances than does conventional OM. However, as we have seen with the low-transition simulations and the two real data sets, when most of the sequences consist of few, long, spells the overall difference is much reduced. In other words, for much typical life course data, the conventional optimal matching algorithm is relatively robust. This is largely because such life course sequences tend to be relatively simple, and certainly are far more simple than truly discrete sequences of similar length.

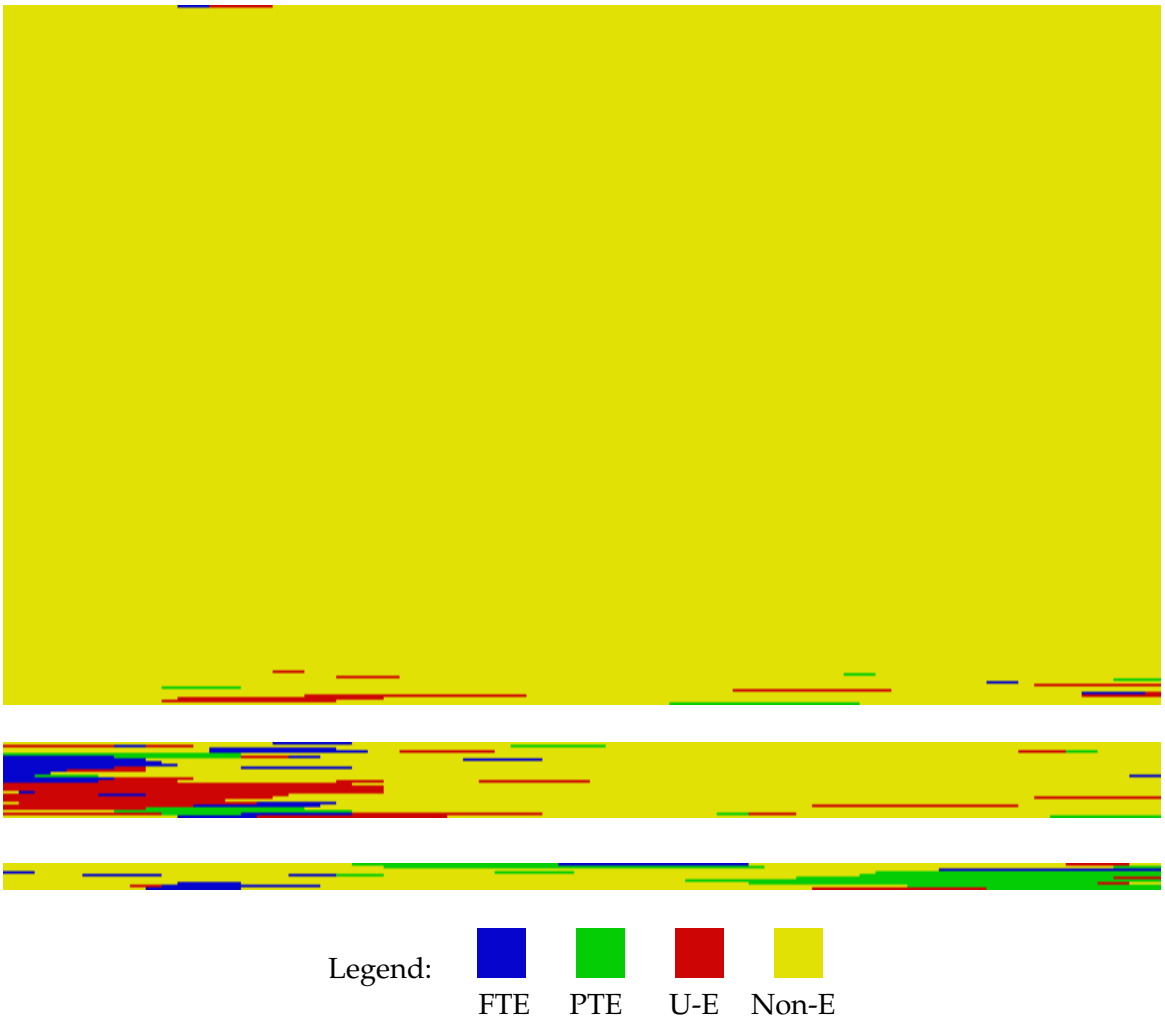

Figure 6: OMv cluster 1 splits in three in OM, respectively clusters 1, 2 and 6

We can take a closer look at the issue of sequence complexity, viewed as entropy or as transition frequency, in Table 4, which reproduces Table 3 but with mean entropy and mean number of spells instead of frequency in the cells. Not only is it the case that the high entropy sequences of the high-transition simulation data set perform differently, but the cases in the BHPS data which cluster differently between the OM and OMv analyses tend to have higher entropy and higher rates of transition. All the substantial off-diagonal cells in the table have substantially higher entropy and transitions than one or both of the corresponding diagonal cells. In other words, the simpler sequences are clustered more stably, but it is mainly the more complex (or chaotic) ones that the two algorithms disagree about. From the point of view of the sociological goal of the analysis, there is good and bad in this: insofar as the goal is to generate a data-driven typology, it is good to find a number of stable clusters containing simpler sequences, but it would also be good to be able to classify the more complex sequences more reliably. It is good to know that a large proportion of mothers are non-employed for the entire duration, and that another large group remain doggedly in full-time work, but it hardly needs sequence analysis to bring this out: as sociologists we are looking for a tool that brings order to the more complex trajectories. That said, in both analyses, the smaller clusters of complex sequences do constitute useful summaries of the structure, albeit ones that are not stable.

Part of the reason for this lack of stability is the nature of cluster analysis. Our goal is to assign

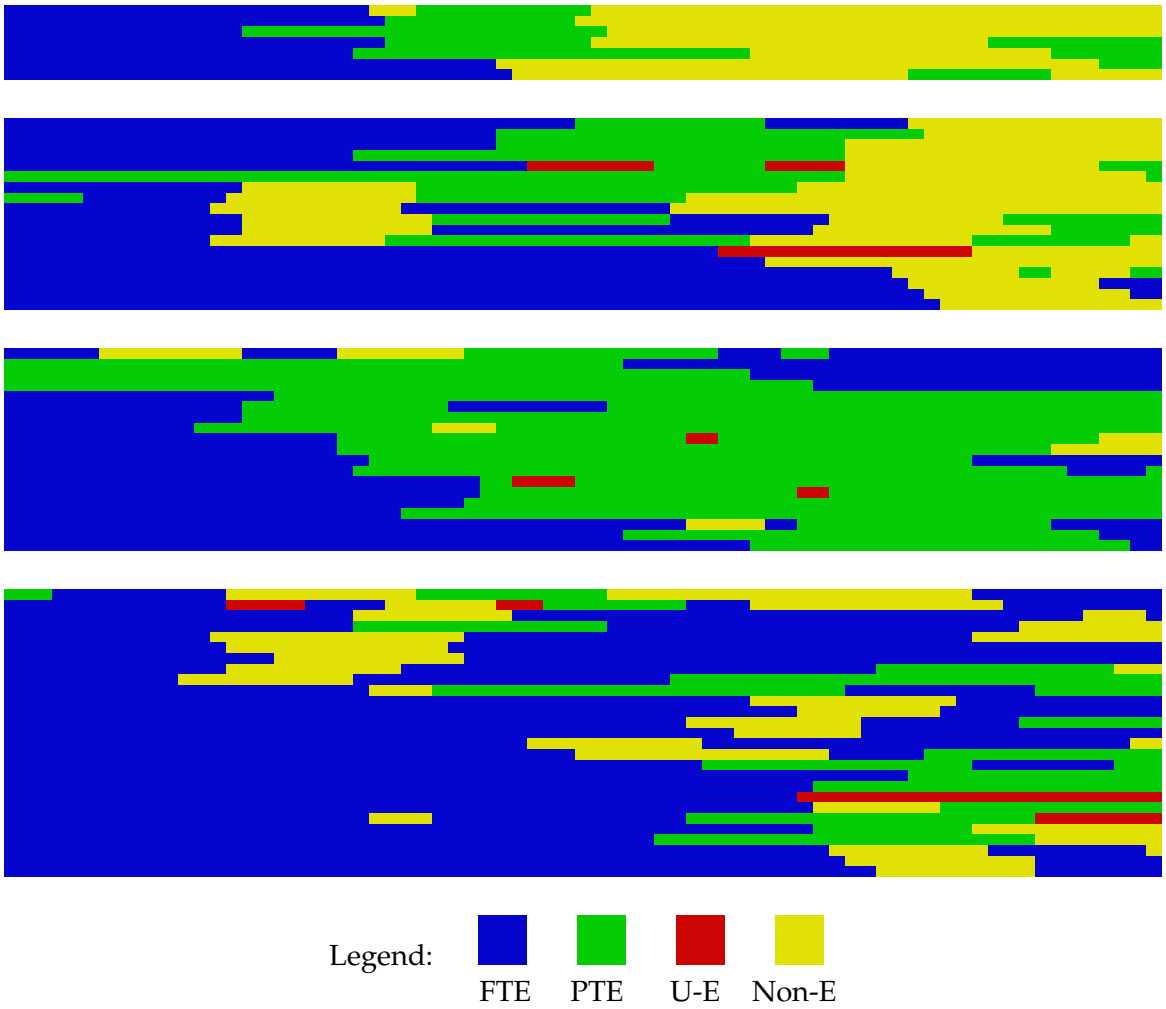

Figure 7: OM cluster 3 splits in four in OMv, respectively clusters 2, 3, 4 and 8

each sequence to one of a small number of categories or clusters. This has heuristic or descriptive value but it is not at all guaranteed that the clusters represent, say, a structure of latent variables. Moreover, given the sort of data we have, the cluster solutions are not likely to be stable at the margins. Multi-dimensional scaling analysis (not shown) makes clear that the inter-sequence distances imply a distribution of the sequences in multi-dimensional space that, while highly structured, is quite even and does not fall into natural clusters.[8]

In this much, perhaps we should explore other manners of post-processing the distances, alternatives to cluster analysis, as much as worry about ways of making the distance measure more sociologically meaningful.

5.1 Time warping and combinatorial methods as alternatives

While the proposed adjustment to the OM algorithm is novel, Abbott and Hrycak (1990, p168ff) discussed the use of transformations of time to take account of other non-linearities, particularly the idea that causal processes may happen at different speeds at different parts of the life course –

[8] A similar analysis of distances from Elzinga's combinatorial method shows much stronger separation of groups, but unfortunately the distinct groups are of sequences which have no elements in common, while the bulk of sequences (and all the more complex ones) are found in one large central group.

Table 4: OM and OMv 8-cluster solutions, entropy and spell-density

OMv	OM							
	1	2	3	4	5	6	7	8
1	0.02	0.90	.	.	.	1.20	0.99	.
	1.11	3.79	.	.	.	4.50	6.00	.
2	.	0.99	1.45	.	1.28	.	.	.
	.	2.79	3.57	.	3.50	.	.	.
3	.	.	1.23	.	.	.	.	.
	.	.	3.89	.	.	.	.	.
4	.	.	1.03	0.04	1.46	.	.	.
	.	.	3.37	1.11	3.00	.	.	.
5	.	.	.	.	1.47	.	.	.
	.	.	.	.	4.24	.	.	.
6	.	.	.	.	.	1.22	.	.
	.	.	.	.	.	3.85	.	.
7	.	.	.	.	.	1.57	0.16	.
	.	.	.	.	.	5.67	1.39	.
8	.	.	1.00	.	.	.	.	0.03
	.	.	3.81	.	.	.	.	1.12

Note: The top figure in each cell is entropy, the lower is mean number of spells. Cells are coloured as in Table 3 – substantial off-diagonal cells are highlighted in red. Entropy is calculated as $-\sum p_i \log_2 p_i$ where p_i is the proportion of months in state i (Wikipedia, 2008).

they proposed using the log of time to generate the discretised sequences to weight later time less. This discussion drew on Kruskal and Liberman (1983), in the context of "time warping". What Abbott and Hrycak proposed was simply a non-linear time-axis, but time-warping in general is rather more powerful, and may present an alternative way of generating more sociologically meaningful distances. Time-warping in this sense is adapted to comparing trajectories in continuous time, by locally compressing or expanding the time axis to make the sequences match. Kruskal and Liberman (1983) describe the method first in continuous terms and then demonstrate that it can be translated to a discrete representation such as monthly histories. This, supplemented by more recent work such as Clote and Straubhaar (2006), who propose a time-warping which is not only symmetric between pairs of sequences but also symmetric in time, and Marteau (2007), who proposes a penalty for the time-warping analogous to indel costs, may provide a starting point for a completely different way of generating sociologically meaningful distances between sequences, that avoids the problem of using methods defined for discrete sequences.

Similarly, Elzinga's combinatorial methods must also be considered, because these, in particular his so-called X/T method where the sequences are composed of spells weighted by their lengths, offer a very attractive logic. The X/T method can readily be extended to weight spells by a non-linear function of lenght (as OMv does), and can also be extended to treat states in the state space as being more or less similar to each other, analogously to substitution costs. With these extensions it may well represent a very serious competitor to optimal matching.

6 Conclusion

This paper has raised the issue of the substantive sociological meaningfulness of the OM algorithm, particularly for continuous-time, episode-structured trajectories such as life course histories. While OM can readily be defended as meaningful for naturally discrete sequences, it does not naturally fit with episode data, and is blind to the distinction between, say, deleting all of a one-month episode and deleting a month from a six-month episode. The OMv algorithm, however, provides a means of calculating distances that reduces the scale of this problem, by weighting the deletion cost inversely with the length of the sequence. The modified algorithm clearly makes a difference to the distances calculated, and makes very large differences to the resulting pairwise distance structure where there is high variability in spell-length. However, perhaps the most interesting finding is that for typical life-course data, with relatively few spells on average, and relatively low variability in the lengths of spells, the modified algorithm makes quite a small difference. In other words, for such data conventional OM is relatively robust to the problems implicit in the discretisation of continuous-time spell data.

Bibliography

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. 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 Hrycak, A. (1990) Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers, *American Journal of Sociology*, **96**(1), 144–85.

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. (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.

Chan, T. W. (1995) Optimal Matching Analysis: A methodological note on studying career mobility, *Work and Occupations*, **22**, 467–490.

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.

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.

Halpin, B. and Chan, T. W. (1998) Class careers as sequences: An optimal matching analysis of work-life histories, *European Sociological Review*, **14**(2).

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.

Kruskal, J. B. and Liberman, M. (1983) The symmetric time-warping problem, in D. Sankoff and J. B. Kruskal (eds), *Time Warps, String Edits and Macromolecules*, Addison-Wesley, Reading, MA, 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.

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 féminins 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*.

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.

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.

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.

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.

Wikipedia (2008) Information entropy, `http://en.wikipedia.org/wiki/Information_entropy`.

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.