Sequence Analysis for Life Course Data

Modifying the OM Algorithm for better duration handling

Brendan Halpin

Dept of Sociology, University of Limerick

12 May 2007, SAI Conference, UL[1]

[1] RCS: omavsai.tex,v 1.3 2007/05/11 16:09:50 brendan Exp brendan

Introduction

- Today I will present a new algorithm for comparing longitudinal trajectory data such as work or life histories
- A modification of the Optimal Matching algorithm, the dominant method in most current sequence analysis
- Very preliminary work, first test of the modified algorithm
- Thanks to comments from Cees Elzinga, and participants in the Geary Institute Seminar, April 2007.

What is Sequence Analysis?

- Sequence Analysis (SA) treats sequences as units, and compares them holistically
- Sequences in this sense are longitudinal structures, *e.g.*,
  - dance (series of steps)
  - conversation (series of utterances)
  - macromolecules like DNA (series of CAGT "bases")
  - life course histories (series of time-units in life course state space)
  - that is, typically linear sequences of observations in a discrete state space
- Note the distinction between discrete sequences in discrete spaces and continuous-time sequences in discrete spaces: the latter we represent as sequences by discrete-ising time

What is Sequence Analysis?

- SA works by defining pairwise distances between sequences according to some metric
- It typically proceeds by using the pairwise distances to generate data-driven typologies using cluster analysis
- Comparing sequences to reference sequences, and multidimensional scaling are also possible
- This holistic approach is an alternative to more conventional techniques which often focus on transition rates:
    - focus on the outcome (epiphenomenon?)
    - rather than the underlying generative processes

Descriptive, exploratory

- Of descriptive, exploratory value rather than analytic/stochastic
- Can provide an digestible overview of complex longitudinal data
- Of particular promise with multi-domain data
- Or where generative processes are complex and changing (*e.g.*, along the lifecourse)
- However, when you want to test clearly specified hypotheses about the generative processes, conventional stochastic techniques such as hazard rate modelling are often much more powerful

Defining pairwise distance

- Defining the pairwise distance is the foundation of SA
- Many possibilities:
  - Hamming: $D_{AB} = \sum_i \ _{Ai} - \ _{Bi}$ or $\sum_i \mathrm{d}(\ _{Ai}, \ _{Bi})$
  - Degenne: $D_{AB} = \sum_i \cos^{-1}(\mathbf{X}_{Ai}, \mathbf{X}_{Bj})$ where $\mathbf{X}_{Ai}$ the cumulated duration
  - Dijkstra–Taris (1995): delete repeats, delete non-common elements, count matches
  - Optimal matching algorithm (OMA): count number of “edits” to change one sequence into another, extensive use in molecular biology
  - Elzinga (2003, 2005): count the number of times the same states occur in the same order in two sequences

Hamming distance

- Hamming distance is very simple but limited
  - `ABABAB` and `BABABA` maximally different – temporal rigidity
  - Less of a problem with life course data, typically has long runs in same state
  - Buchmann and Sacchi (1995) factor analyse occupational characteristics to define inter-occupational distances
  - Lesnard (2006) proposes a "dynamic" Hamming distance for time-diary data: distances defined by transition rates between activities, varying throughout the day
  - If time has a meaningful ruler (as with daily time use) Hamming's time-rigidity is a positive advantage
  - For more elastic "developmental" time, perhaps less appropriate

Degenne to Elzinga

- Degenne’s method promising but untried
- Dijkstra–Taris more or less directly superseded by OMA
- Elzinga’s method shows a lot of promise, plus has strong intuitive basis

The Optimal Matching Algorithm–detail

- Given two sequences, s_1 and s_2, drawn from an "alphabet" S, the Optimal Matching Algorithm generates a distance measure based on "elementary operations":
  - substitution and
  - insertion and deletion
- Substitution replaces an element in s_1 with the corresponding element in s_2
- Insertions and deletions equivalently delete an element in s_1 or insert an element in s_2 (or vice versa); because of the equivalence they are known as *indel*s

The Optimal Matching Algorithm–distance

- The distance between s_1 and s_2 is defined as the least expensive path from one to the other using the elementary operations
- OMA allows a matrix of costs for all pairwise substitutions, and an indel cost to be specified
- The OM algorithm is a dynamic programming technique which finds the “cheapest” path in a time- and memory-efficient manner

Abbott’s evangelism

- Andrew Abbott has been the main evangelist for OMA in sociology
- A string of articles from Abbott and Forrest (1986), Abbott and Hrycak (1990) *etc.*to a retrospective review (Abbott and Tsay, 2000) with a related debate (Levine, 2000; Wu, 2000; Abbott, 2000)
- In Abbott (1995a), demonstrates OMA to be more general than Dijkstra–Taris

OM in social science literature

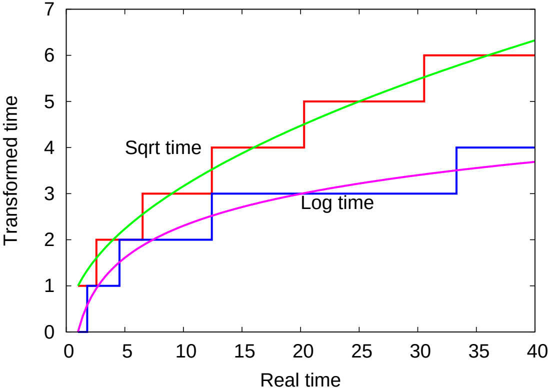
- OM is becoming widely used, particularly since Götz Rohwer incorporated it in TDA
    - Class careers: Halpin and Chan (1998)
    - Women's careers in finance: Blair-Loy (1999)
    - Transition from school to work: Scherer (2001), McVicar and Anyadike-Danes (2002)
    - Methods paper: Brüderl and Scherer (2004)
    - Male careers: Anyadike-Danes and McVicar (2005)
    - Time use: Wilson (2006)
    - Gendered careers: Levy et al. (2006)
    - Housing, employment, marriage, fertility: Pollock (2007)
- More recently, Kohler et al have released a Stata package, `SQ`, for OMA

Problems with OM for lifecourse data

- For sequences that are naturally discrete in time, OM works well
- But life course sequences are have highly variable spell lengths, and the discrete representation may not suit OM as well
- For instance, given $s_1 = \texttt{ABBD}, s_2 = \texttt{ABCD}, s_3 = \texttt{ABDD}$, all three sequences will be equidistant
- But sociologically, s_1 and s_3 are clearly closer
- OM doesn't recognise the continuity; a slight adjustment of spell length is treated as being as expensive as the introduction of another spell

Change the time scale?

- One suggestion (Abbott made it in early work) is to change the time scale: use log time, for instance
- However, this exacerbates the discretisation



Modifying the OM algorithm

- I propose instead a modification of the OM algorithm that has an analogous effect: scale the costs of elementary operations according to the length of the affected spells
- There are limits to what we can change in the algorithm without degrading its performance
- In particular, the algorithm has no memory (*e.g.*, "now deleting an element from this spell for the second time")
- However, we can take account of spell length in setting costs

The OM algorithm in detail

- I begin by outlining the operation of the OM algorithm
- It uses dynamic programming techniques to efficiently determine the cheapest set of edits to transform one sequence into another – hence “optimal”
- Operates by calculating the elements of a matrix where each element $C_{ij} = min(c_{i-1,j-1} + !\ _{i,j}, c_{i,j-1} + \ \ , c_{i-1,j} + \ \ )$,
- ! is the substitution matrix, and the first row and column are filled with the cumulative insertion/deletion costs
- A diagonal move represents a substition, right represents deletion, down represents insertion
- Bottom right cell eventually contains the optimal cost

Working through OM

Cell value: $min(c_{i-1,j-1} + !\ \ _{i,j}, c_{i,j-1} + \ \ , c_{i-1,j} + \ \ )$

$= min(0+2, 2+2, 2+2) = 2$

$= min(2+1, 2+2, 4+2) = 3$

$= min(4+0, 3+2, 6+2) = 4$

$= min(6+1, 4+2, 8+2) = 6$

s_1

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

0	2	4	6	8
2	2	3	4	6
4	4	4	4	4
6	4	5	6	6
8	6	5	7	8
10	8	6	6	8

Tracing the operations

To convert ABCD into CDAAB the following set of operations gives the cheapest path:

Operation	Intermediate state	Cost
	ABCD	= 0
insert C	CABCD	+2 = 2
insert D	CDABCD	+2 = 4
const A = A	CDABCD	+0 = 4
subs B→A	CDAACD	+1 = 5
subs C→B	CDAABD	+1 = 6
delete D	CDAB–	+2 = 8

Where sequence lengths are variable the distance is often scaled inversely with the longer of the pair: 8 units thus become an pairwise distance of 1.6

Efficient

- This strategy calculates the optimum path efficiently
- Memory requirements are proportional to $l_1 \times l_2$
- Time requirement is $a + b(l_1 \times l_2)$
- Both are small relative to the $O(N^2)$ required to process every pairwise comparison between N sequences
- Nonetheless, for realistic data sets the procedure completes acceptably fast
- **However**, adaptations of the algorithm risk destroying its efficiency, for instance by requiring exponentially growing memory or processing time

Variant OMA

- The modified algorithm I propose treats spell length in the following manner
  - single unit spell costed as per OM
  - total cost of a multiple unit spell to be strictly increasing in length
  - cost of each unit to be strictly decreasing with length of spell
- Costing units at $basecost \times \frac{1}{\sqrt{len}}$ achieves this:
  - Cost of 1 unit spell: $basecost \times 1$
  - Cost of 2 unit spell: $basecost \times (0.707 + 0.707) = basecost \times 1.414$, *etc.*
- So will other exponents, other functions

Indel and substitution

- Insertion and deletion costs are modified directly and equally, as they are equivalent: an insertion in s_1 is equivalent to a deletion in s_2
- Substitution costs must also be modified, as a substitution is a deletion followed by an insertion
- The direction of the substitution is unimportant, so we cost it as if it involved a deletion in the longer subsequence
- If s_{1i} is unique, and s_{2j} is part of a run of two, the substitution cost is divided by $\sqrt{2}$
- If s_{1i} is one of three and s_{2j} one of two, division is by $\sqrt{3}$

The matrix operations

$$
\begin{array}{cc|c|c|c|c|}
 & & \multicolumn{4}{c}{s_1} \\
 & & A & B & C & D \\
\cline{3-6}
 & C & 2 & 1 & 0 & 1 \\
\cline{3-6}
s_2 & D & 3 & 2 & 1 & 0 \\
\cline{3-6}
 & A & 0 & 0.7 & 1.4 & 2.1 \\
\cline{3-6}
 & A & 0 & 0.7 & 1.4 & 2.1 \\
\cline{3-6}
 & B & 1 & 0 & 1 & 2 \\
\cline{3-6}
\end{array}
\qquad
\begin{array}{c|c|c|c|c|}
0 & 2 & 4 & 6 & 8 \\
\cline{2-5}
2.0 & 2.0 & 3.0 & 4.0 & 6.0 \\
\cline{2-5}
4.0 & 4.0 & 4.0 & 4.0 & 4.0 \\
\cline{2-5}
5.4 & 4.0 & 4.7 & 5.4 & 5.4 \\
\cline{2-5}
6.8 & 5.4 & 4.7 & 6.1 & 6.8 \\
\cline{2-5}
8.8 & 7.4 & 5.4 & 5.7 & 7.7 \\
\cline{2-5}
\end{array}
$$

Typical sequences compared

Sequences		OMA	New variant	
A	B		exp = 0.5	exp = 0.75
1 2 3 4 3	2 3 4 1 4	1.00	1.00	1.00
1 2 3 4 3	2 2 2 1 3	1.00	0.83	0.78
1 2 3 4 3	2 2 2 2 2	1.00	0.45	0.30
2 3 4 1 4	2 2 2 1 3	0.80	0.55	0.46
2 3 4 1 4	2 2 2 2 2	1.20	0.54	0.36
2 2 2 1 3	2 2 2 2 2	0.40	0.18	0.12

Implementation

- Implemented in Stata
- C plugin, platform dependent, fast, relatively simple code
- Platform independent Mata implementation also possible, but very slow by comparison

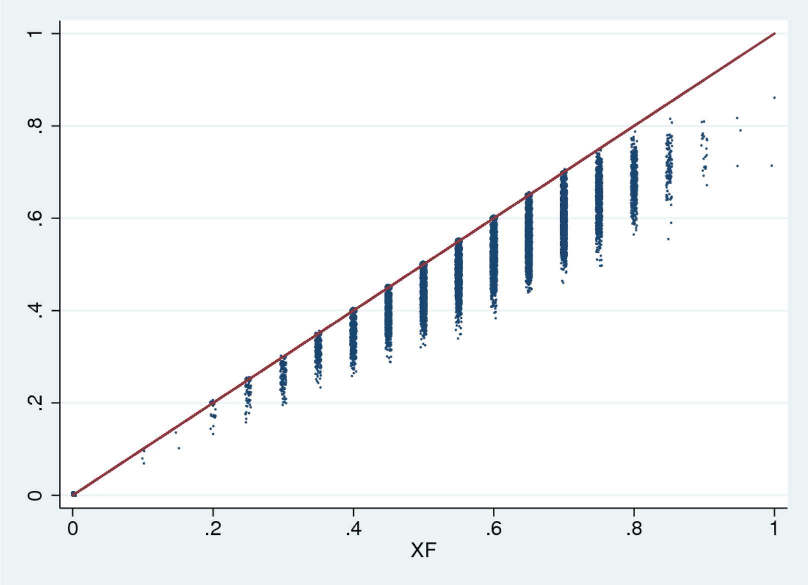
Simulated data

- Tests are run with simulated data, 4 states, 20 time-units, with
  - a low probability of multiple observations in sequence (82% of cases have max run of ≤ 2)
  - a medium probability (79% ≤ 5)
  - and a higher probability (79% ≤ 11)
- As expected, the higher the presence of runs the lower the mean scores
- But correlations remain high, even excluding matches between identical sequences (distance zero)

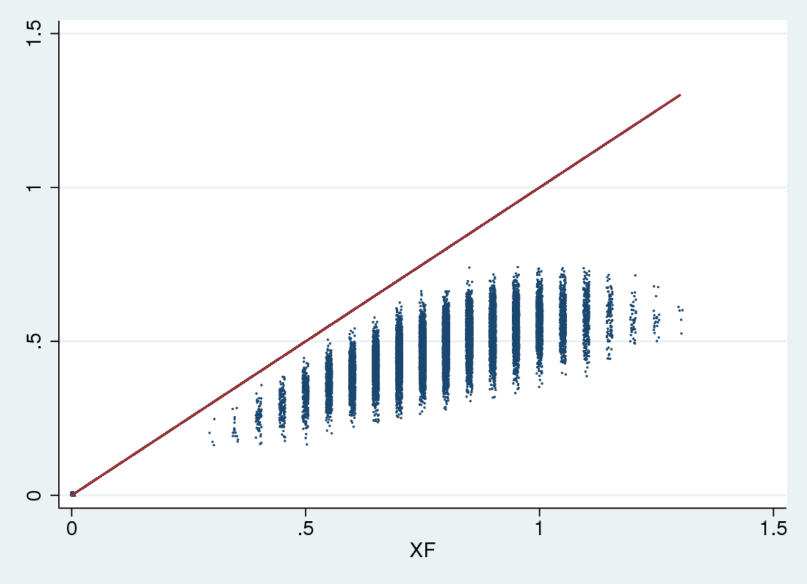
Correlations

		All	Non-identical
Low	Pearson	0.9978	0.9496
	Spearman	0.9984	0.9420
Med	Pearson	0.9931	0.8092
	Spearman	0.9946	0.7985
High	Pearson	0.9821	0.8128
	Spearman	0.9949	0.8132

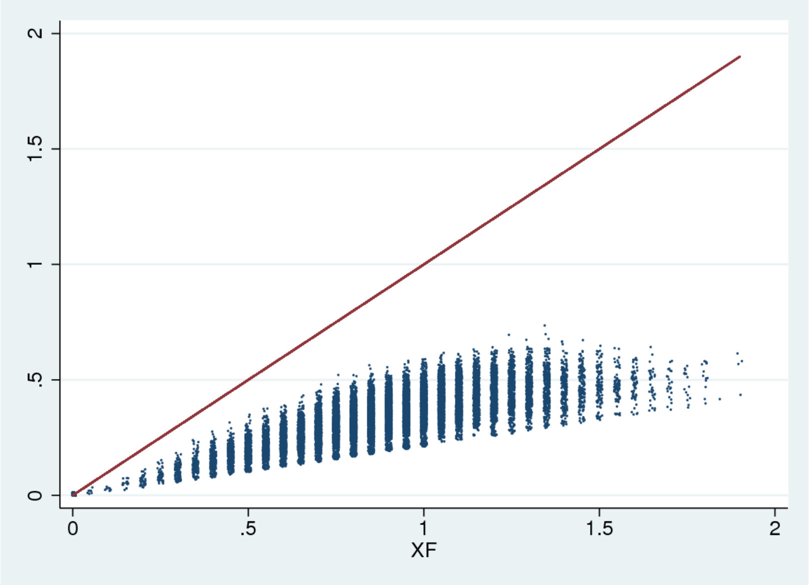
Correlation of OMA and OMAv: low level of runs



Correlation of OMA and OMAv: medium level of runs



Correlation of OMA and OMAv: high level of runs



Empirical typologies

- But what about real data and the "end product"?
- Typically "empirical typologies" generated by cluster analysis
- Using BHPS fertility and labour market histories, I construct 5-year labour market histories for women: 2 years before and 3 years after a birth
- Encoded as monthly status, 4 states:
    - Full-time employed
    - Part-time employed
    - Unemployed
    - Not in labour market

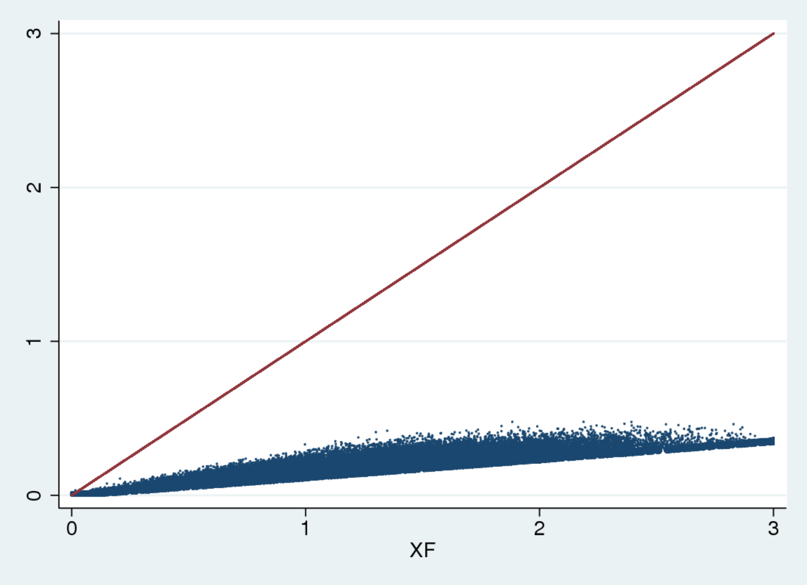
OM costings

- Substitution matrix:

FTE	0	1	2	3
PTE	1	0	1	2
UE	2	1	0	1
NonE	3	2	1	0

- *Indel* cost: 2 units

Correlation of OMA and OMAv: labour market data



Results

- Correlation of 0.97 between OMA and OMAv, excluding identical sequences – higher than expected
- Spearman correlation similar

Cluster solution

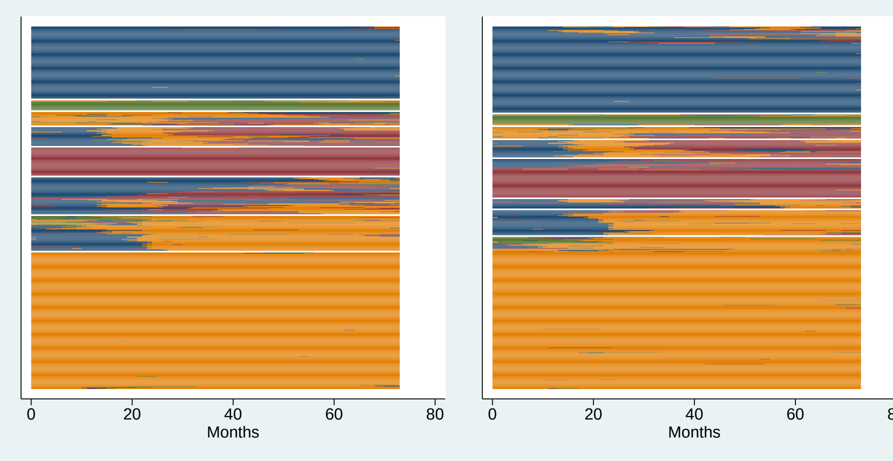
- Fit an 8 cluster solution to both data sets
- By inspection gives an acceptable result
- Very high level of agreement, especially for “low entropy” sequences (*i.e.*, near 100% dominated by a single state)
- Less agreement where more is “going on”

Comparing the eight-cluster solution

OMAv	OMA								Total
	1	2	3	4	5	6	7	8	
1 (slide 32)	263	27	0	0	0	2	1	0	293
2 (slide 33)	0	39	7	0	2	0	0	0	48
3 (slide 34)	0	0	18	0	0	0	0	0	18
4 (slide 35)	0	0	19	54	1	0	0	0	74
5 (slide 36)	0	0	0	0	33	0	0	0	33
6 (slide 37)	0	1	0	0	0	21	0	0	22
7 (slide 38)	0	0	0	0	0	3	18	0	21
8 (slide 39)	0	0	27	0	0	0	0	139	166
Total	263	67	71	54	36	26	19	139	675

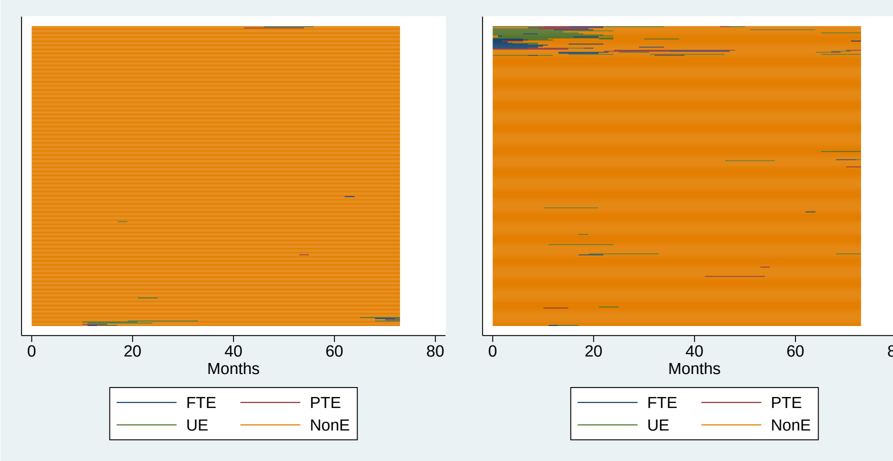
(41,42)

8 cluster solution – OMA (left) and OMAv (right)



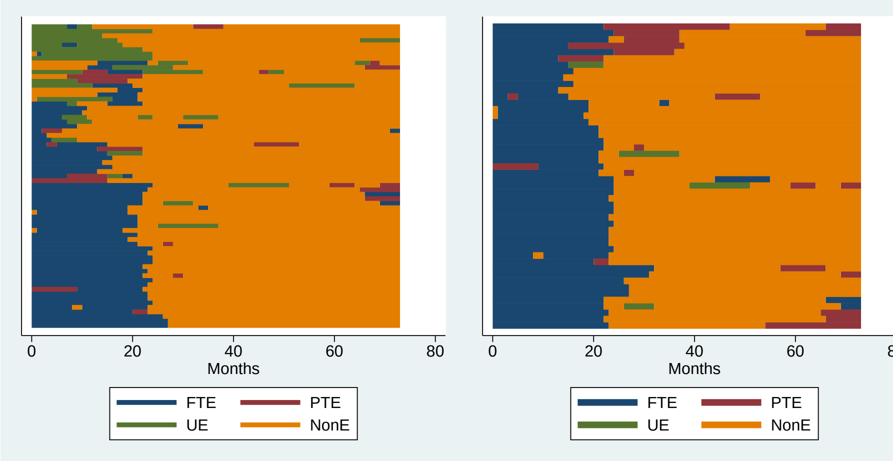
(Table: slide 30)

Cluster 1 – OMA (left) and OMAv (right)



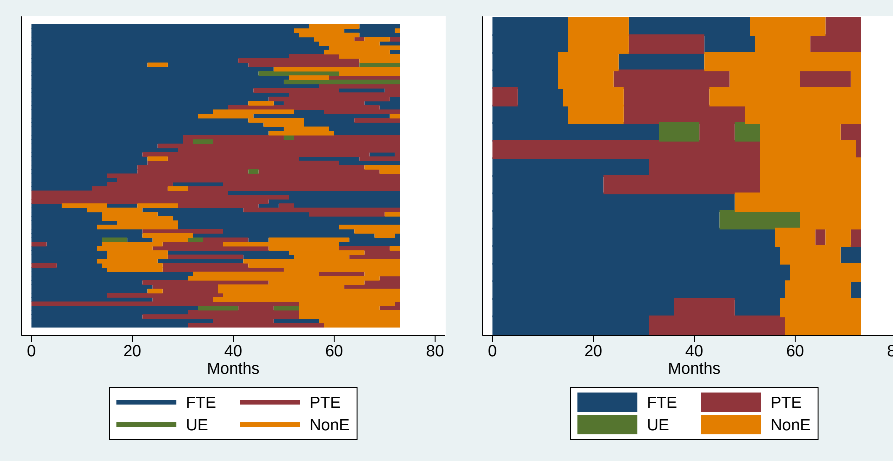
(Table: slide 30)

Cluster 2 – OMA (left) and OMAv (right)



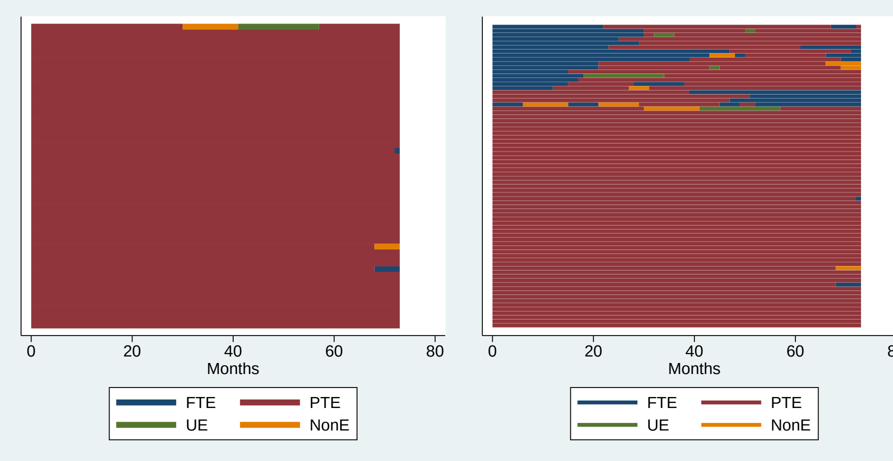
(Table: slide 30)

Cluster 3 – OMA (left) and OMAv (right)



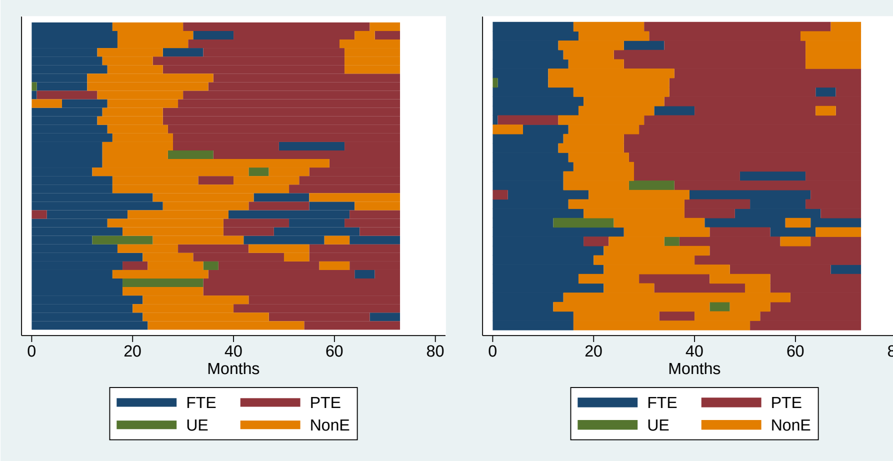
(Table: slide 30)

Cluster 4 – OMA (left) and OMAv (right)



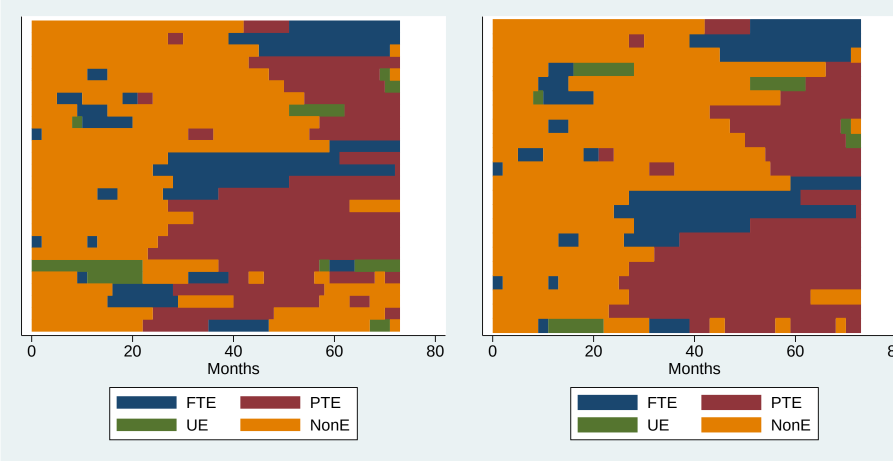
(Table: slide 30)

Cluster 5 – OMA (left) and OMAv (right)



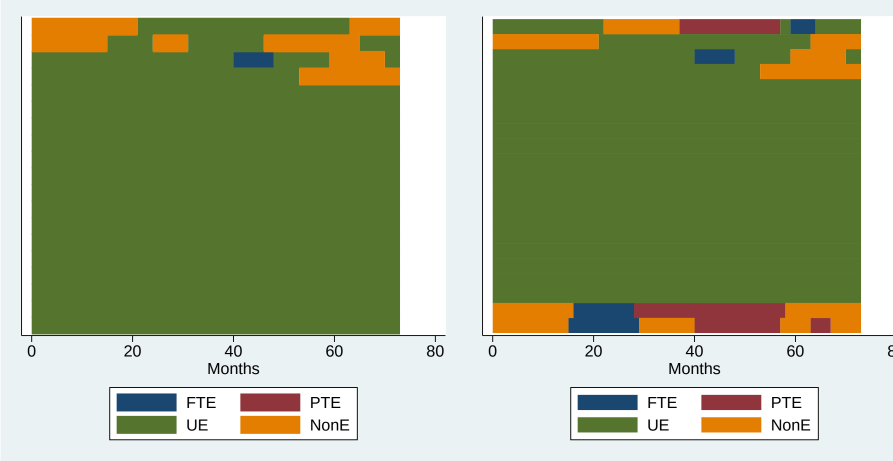
(Table: slide 30)

Cluster 6 – OMA (left) and OMAv (right)



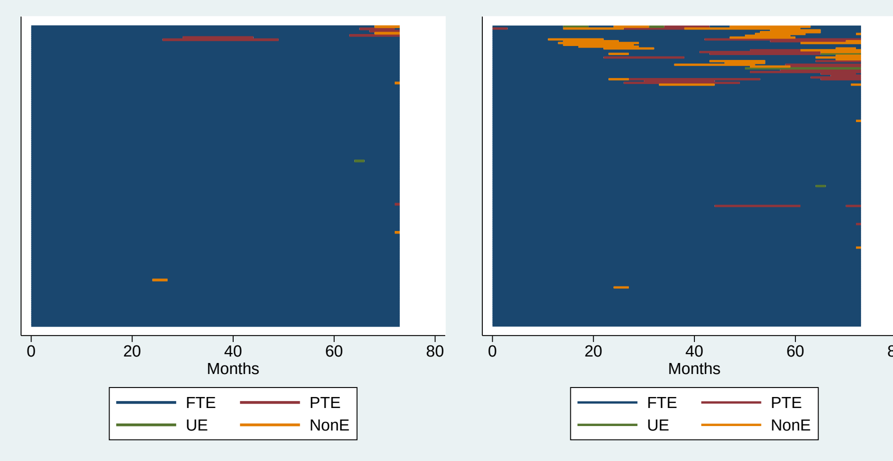
(Table: slide 30)

Cluster 7 – OMA (left) and OMAv (right)



(Table: slide 30)

Cluster 8 – OMA (left) and OMAv (right)

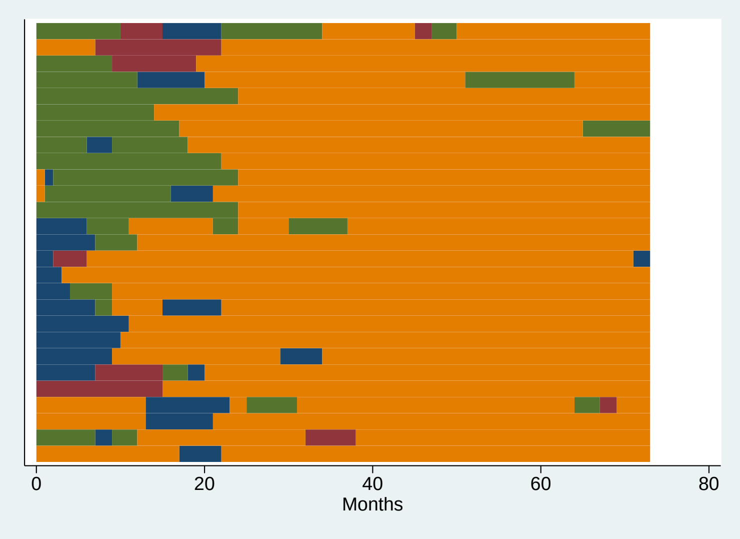


(Table: slide 30)

Comparing the clusters

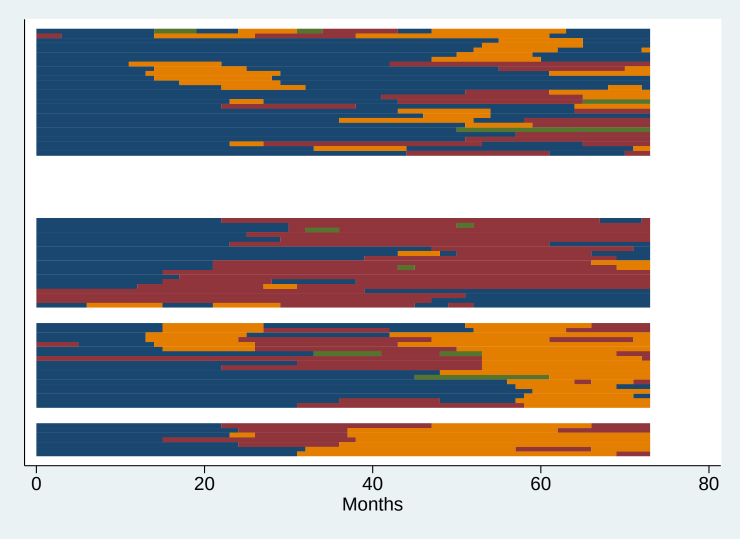
- High agreement
- Two major sources of disagreement
  - 27 cases move from OMA cluster 2 to OMAv cluster 1
  - OMA cluster 3 scattered across OMAv clusters 2, 3, 4 and 8
- The 27 OMA cluster 2 cases are arguably better off in cluster 1 (dominated by non-employment) than cluster 2, where early unemployment is matched to early FT-employment

OMA Cluster 1/2 split



(Table: slide 30)

OMA Cluster 3 split



(Table: slide 30)

OMA Cluster 3 details

- Cases moved to cluster 8 are initially FTE with lots of transitions in and out of employment later, and are moved to the predominant FTE cluster
- Those moved to cluster 4 transition from FTE to PTE around the birth, and largely stay there, and merge with the predominant PTE cluster
- Those that remain in cluster 3 are a distinct group: initially FTE, try to remain in the labour market but finally drop out
- to cluster 4: FTE with late shift to PTE, matched to predominant FTE cluster
- Those that move to cluster 2 are also initially FTE but drop out, typically with a short spell of PTE, and are matched with very similar trajectories that do not have the short PTE spell

Conclusions

- OMAv and OMV generate very but not completely similar results
- To the extent that they differ, it is arguable that it gives a superior clustering of the more complicated trajectories
- Extent of similarity is greater than I had expected
  - Conventional OMA may be adequate where all sequences characterised by long runs
  - Greater mix of long and short runs may be different – result is likely to depend on the data to some degree
- As seen, the differences are greatest in the clusters characterised by higher "entropy" sequences – these are the ones most sensitive to the distance measure; even naïve matching (*e.g.*, Hamming) will match the simple cases
- "More research is needed"

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

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

Abbott, A. (1995a). A comment on "Measuring the agreement between sequences". *Sociological Methods and Research*, 24(2):232–243.

Abbott, A. (1995b). 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 Forrest, J. (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History*, XVI(3):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):144–85.

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.

Bradley, D. W. and Bradley, R. A. (1983). Application of sequence comparison to the study of bird songs. In Sankoff and Kruskal (1983), chapter 6.

Brüderl, J. and Scherer, S. (2004). Methoden zur analyse von sequenzdata. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 44:330–347.

Buchmann, M. and Sacchi, S. (1995). Mehrdimensionale Klassifikation beruflicher Verlaufsdaten: Eine Anwendung auf Berufslaufbahnen zweier Schweizer Geburtskohorten. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 47(3):413–442.

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

Degenne, A., Lebeaux, M.-O., and Mounier, L. (1996). Typologies d'itinéraires comme instrument d'analyse du marché du travail. Troisièmes journées d'études Céreq-Cérétim-Lasmas IdL, Rennes, 23–24 May 1996.

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

Elzinga, C. (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. (2003). Tracks through time and continuous processes: Transitions, sequences, and social structure. Working paper 2003-01, Dept of Sociology, University of Limerick.

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

Kruskal, J. B. (1983). An overview of sequence comparison. In Sankoff and Kruskal (1983).

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.

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.

Sankoff, D. and Kruskal, J. B., editors (1983). *Time Warps, String Edits and Macromolecules*. Addison-Wesley, Reading, MA.

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

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.