Simulating Sequences

Brendan Halpin
Department of Sociology
University of Limerick[1]

Journée Trajectoires '11, October 14 2011
Université Paris I

[1]Work in progress – brendan.halpin@ul.ie

Sequence analysis and lifecourse data

- What does SA do for us with lifecourse data?
- Life course data: long spells, few states, important individual characteristics and an "interesting" time dimension
- What does clustering pairwise inter-sequence do for us?
  - Descriptive overview, visualisation – enough?
  - Can it pick up things other techniques miss?
- Today I discuss using simulations, both "artificial" and data-based, to address this question

Brendan Halpin
Department of Sociology
University of Limerick

Sequences are messy

- Lifecourse sequences are epiphenomena of more fundamental underlying processes
- The processes are potentially complex: difficult to predict distribution of sequences
- Other techniques (hazard rate models, models of late outcome using history, models of the pattern of transition rates) give a powerful but partial view
- SA clearly allows us visualise complex data; possibly allows us observe features that will otherwise be missed

Brendan Halpin
Department of Sociology
University of Limerick

Processes

- The generating processes are complex:
  - individuals bring different characteristics from the beginning
  - history matters, including via duration dependence (individuals accumulate characteristics)
  - time matters:
    - calendar time (e.g. economic cycle), state distribution may change dramatically
    - developmental time (maturation)
    - processes in other lifecourse domains
- Too many parameters to model, hard to visualise distribution of life courses, also the possibility of *emergent* features

Brendan Halpin
Department of Sociology
University of Limerick

Outline of the presentation

- Use of "ideal" simulations to test how well SA can recover information about the generative processes
- Looking at an alternative visualisation of sequence structure: a time-dependent average transition matrix
- Using this structure to create data-based simulations, to ask more precise questions of the sequence analysis, including some simple hypothesis testing

Brendan Halpin
Department of Sociology
University of Limerick

Simple simulation

- The purpose of the first exercise is to simulate sets of sequences with a very simple, known structure, and examine how our usual practice (cluster analysis of pairwise OM distances) can recover this structure
- A 3-state space, sequences 40 units long
- Four simple scenarios:
  - Two distinct transition matrices, constant over time
  - One transition matrix, but two different rates of transition
  - One matrix, subgroup is initially faster, then slower
  - One matrix, subgroup is forced to state 3 at a random point
- Unrealistic, little structure, no history, time almost absent

Brendan Halpin
Department of Sociology
University of Limerick

Simulation process

- The test is whether the cluster solution is associated with the generating type
- Association is a much weaker requirement than actually recovering the type information
- 1000 sequences generated at a time, different cluster solutions considered, underlying rate of transition varied

Brendan Halpin
Department of Sociology
University of Limerick

Summary result

Mean p-value of χ^2 test, 16-cluster solution

Monthly transtion rate	Different matrices	Forced state at random point	Different rates of transi-tion	Changing rates of transi-tion
2%	0.037	0.000	0.048	0.041
5%	0.055	0.000	0.026	0.011
10%	0.090		0.023	
15%	0.107			
18%	0.133			

These distributions are skewed, so the proportion significant is higher than the average would suggest. The 0.037 in the first simulation corresponds with 60% of cases with $p < 0.01$.

Brendan Halpin
Department of Sociology
University of Limerick

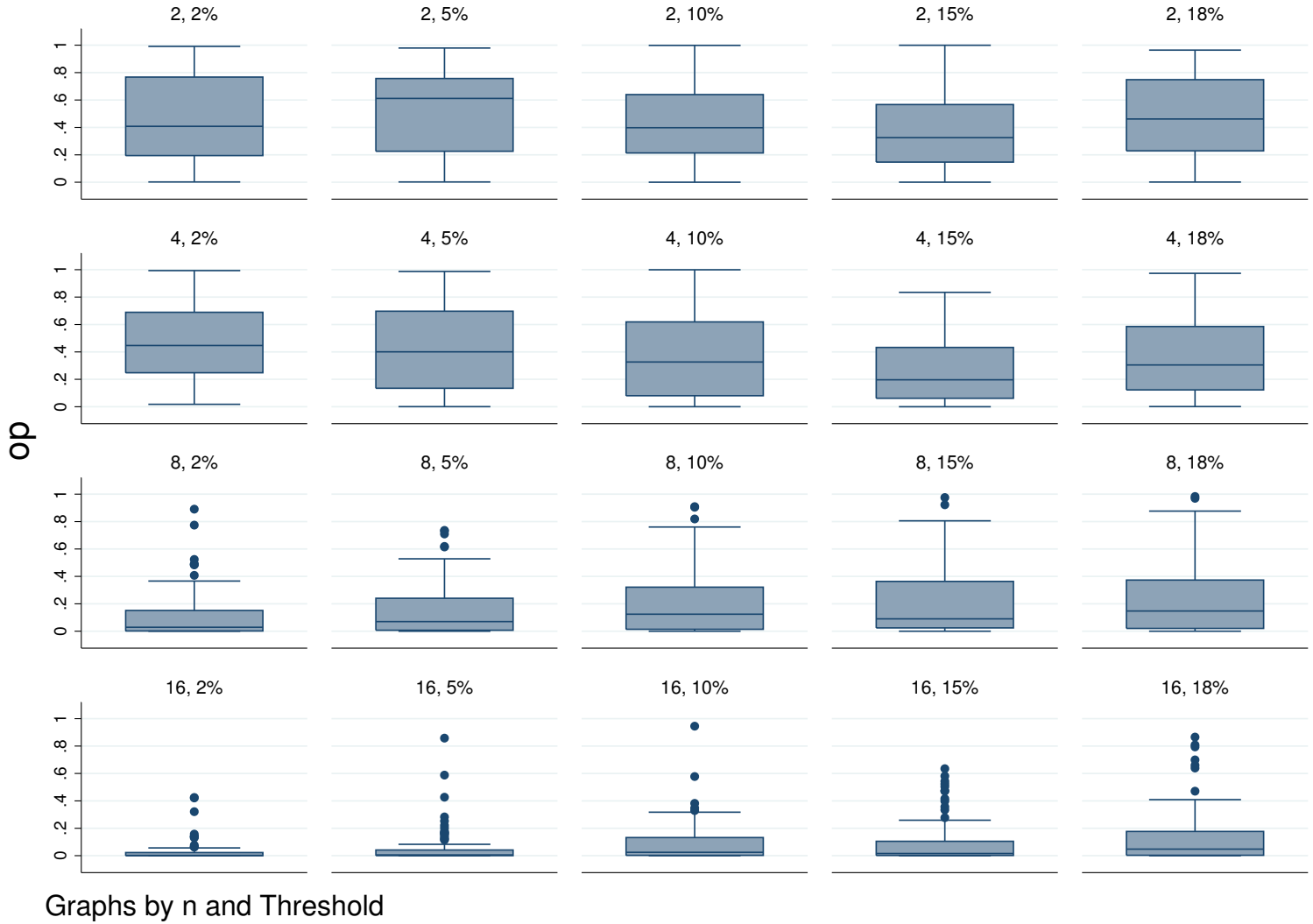
More detail: scenario 1

- Sequences are assigned to a random starting point
- Two transition regimes:

	Type 1			Type 2		
	A	B	C	A	B	C
A	0.80	0.10	0.10	0.80	0.16	0.04
B	0.10	0.80	0.10	0.04	0.80	0.16
C	0.10	0.10	0.80	0.16	0.04	0.80

- State distribution won't change, is the same across type, mean number of spells is the same across type
- "Threshold" parameter to vary monthly rate of change
- Type 2 will simply generate more A, B, C sequences etc., than type 1

Distribution of χ^2 p-values, by rate of transition and cluster size



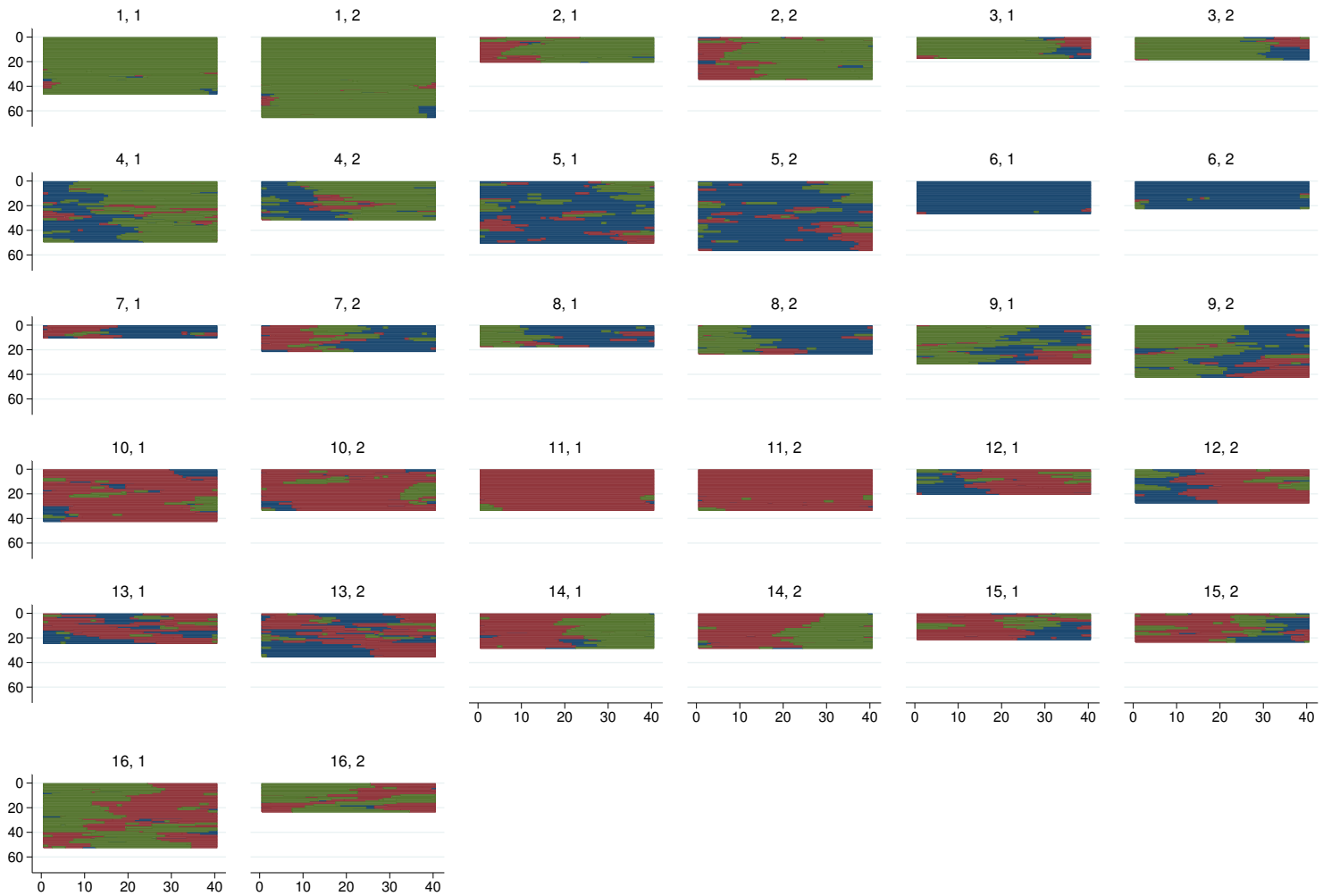
Brendan Halpin
Department of Sociology
University of Limerick

Association, not recovery of pattern

- We that for the 16-cluster solution we have a very good chance of detecting the structure
- However, we can't "recover" the structure so easily
- Inspection of the cluster solution is interesting
- But even enumerating sequences according to their structure (e.g., counts of AB, BC and CA transitions) will only correctly identify about 60% of them

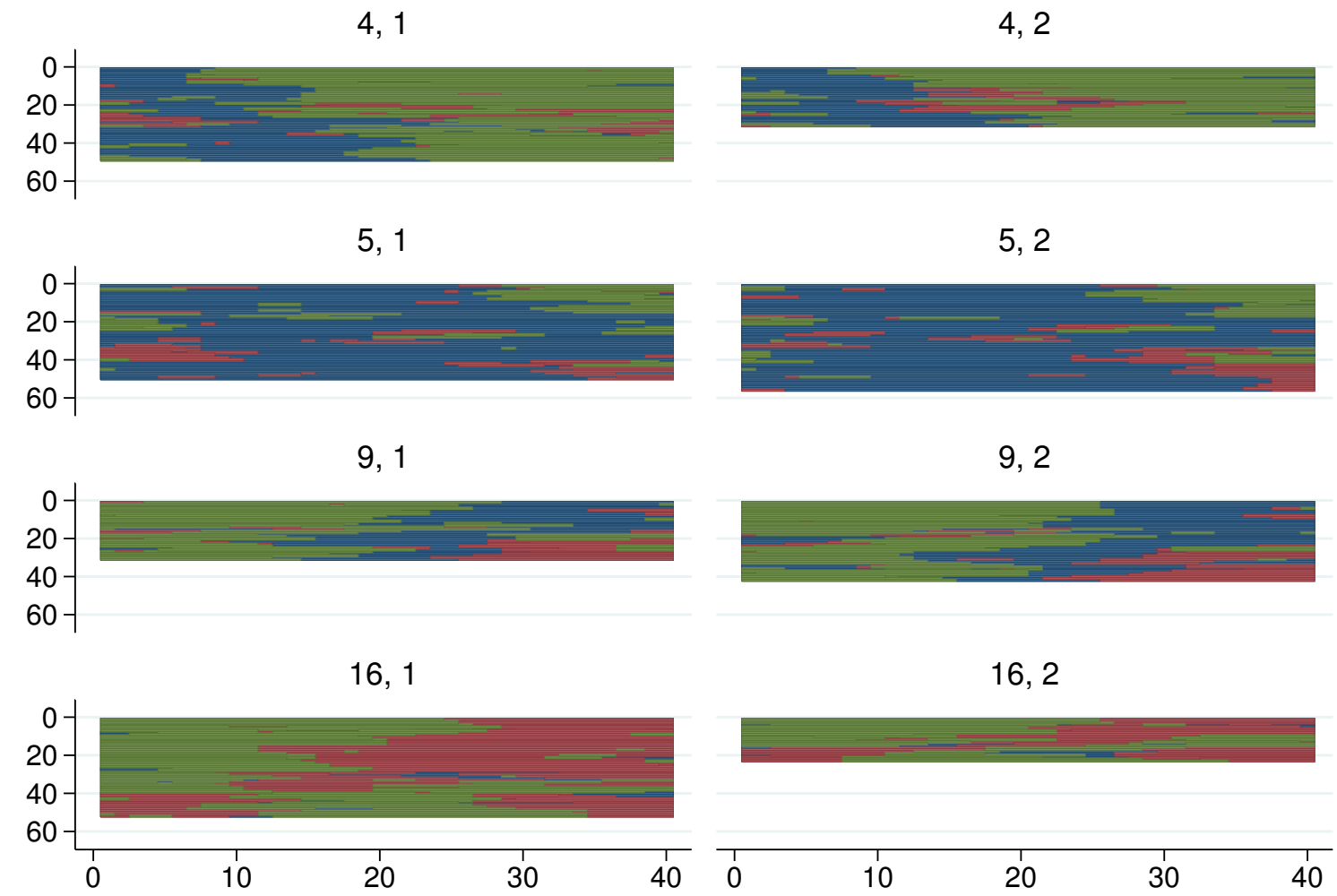
Brendan Halpin
Department of Sociology
University of Limerick

Cluster solution



Graphs by g16 and type

Cluster solution: some key groups

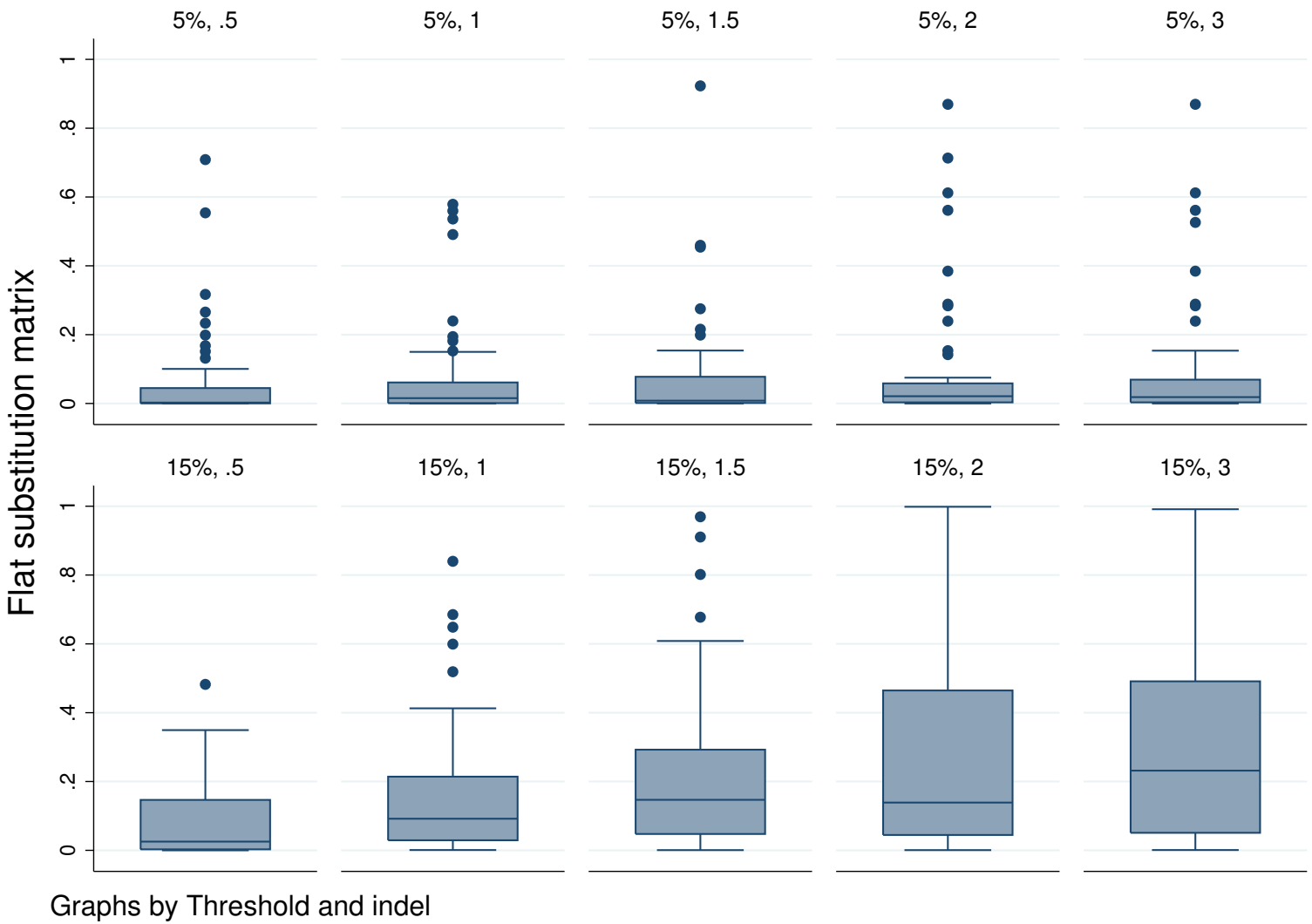


Brendan Halpin
Department of Sociology
University of Limerick

Parameterising OM

- We can use this framework to explore the effect of indel and substitution costs
- Indel costs can vary from half the max substitution cost up – low values make it easy to detect similarity at different times, higher values reduce OM to Hamming distance
- The base simulation uses a substitution cost based on the three states being equally different ("flat" or 2-D solution)
- What if $A \rightarrow C$ is twice $A \rightarrow B$ ("linear" or 1-D solution)

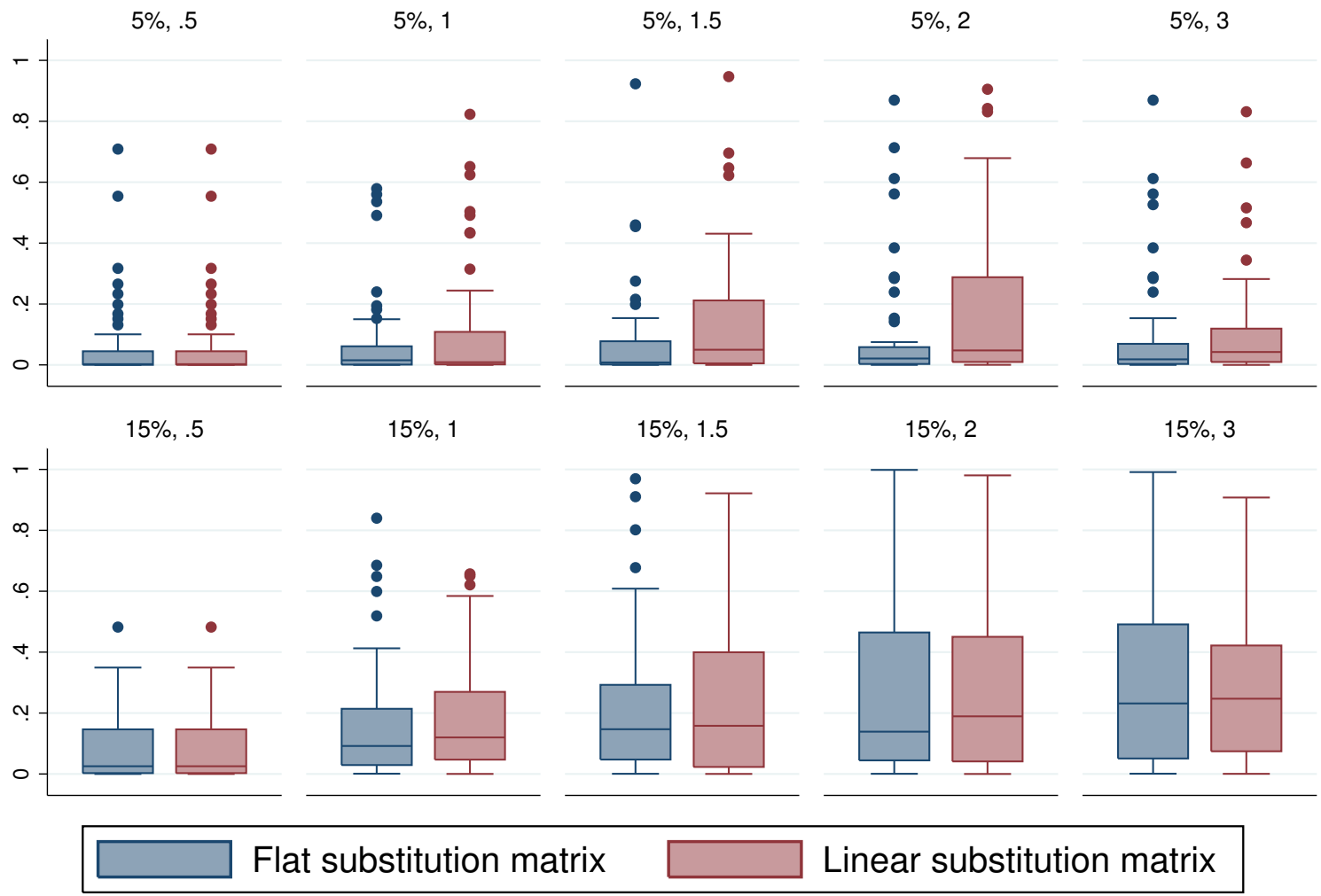
Parameterising OM: Varying indel costs



(800 sequences, 50 replications)

Brendan Halpin
Department of Sociology
University of Limerick

Parameterising OM: Varying substitution costs



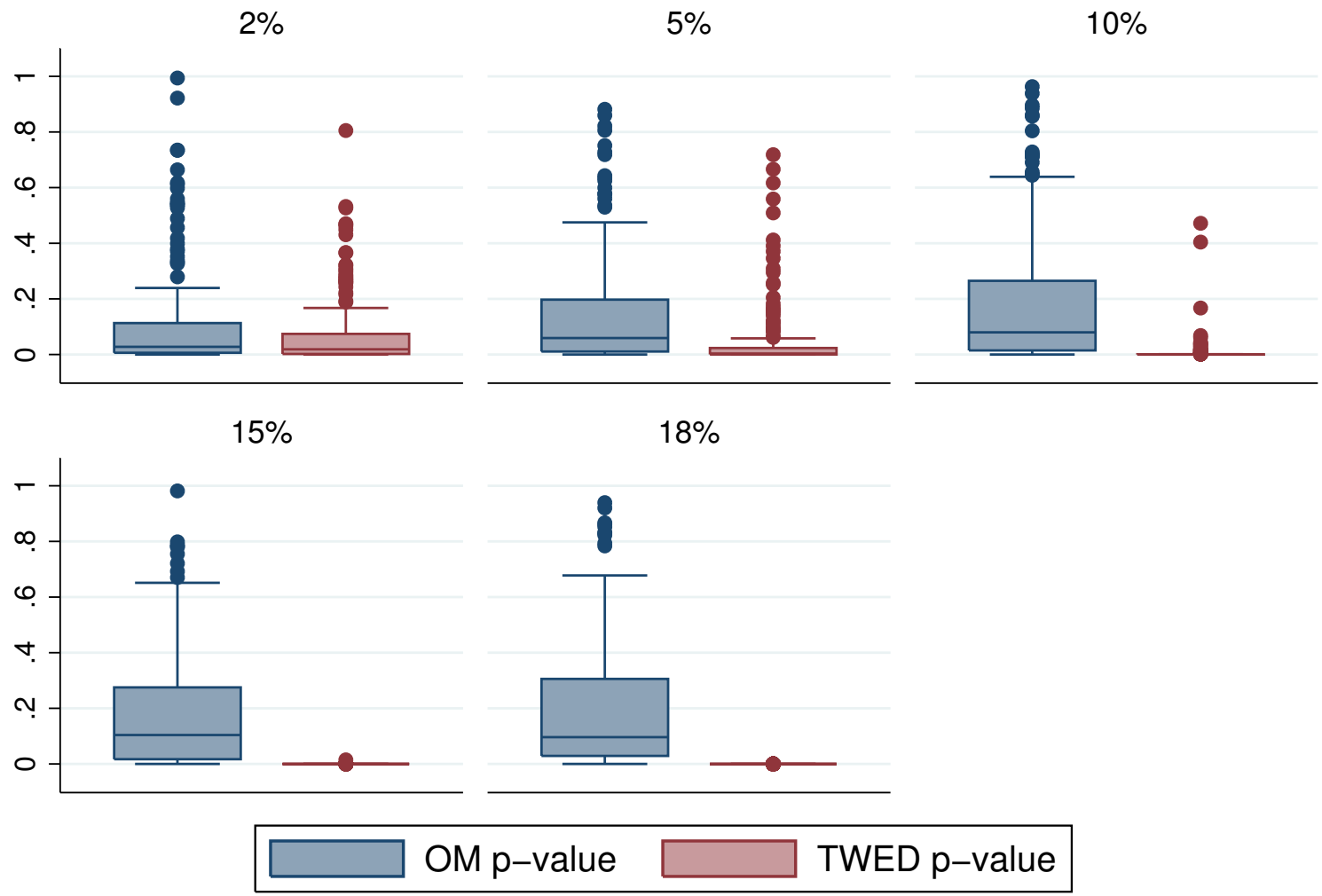
(1000 sequences, 50 replications)

Time-warping

- An alternative distance measure to OM
- Local expansion and compression of the time dimension
- Subtly different logic, similar to implement, somewhat different results
- Described (as TWED) by Marteau (2007, 2008), implemented as a Stata plugin
- See http://teaching.sociology.ul.ie/seqanal/naplestw.pdf for more info
- What difference does it make here?

Brendan Halpin
Department of Sociology
University of Limerick

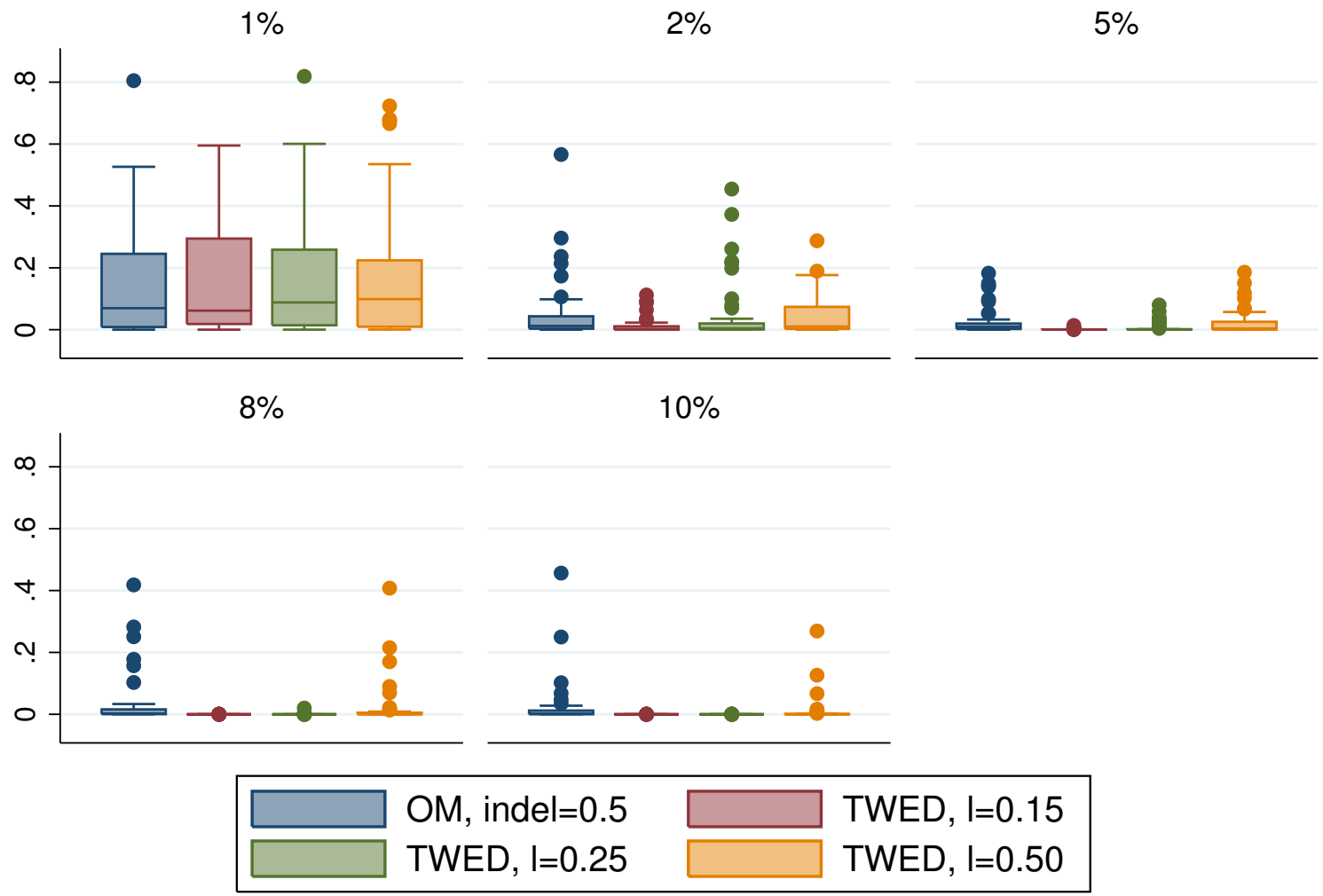
TWED and OM in 2-matrix scenario



600 sequences, 200 replications

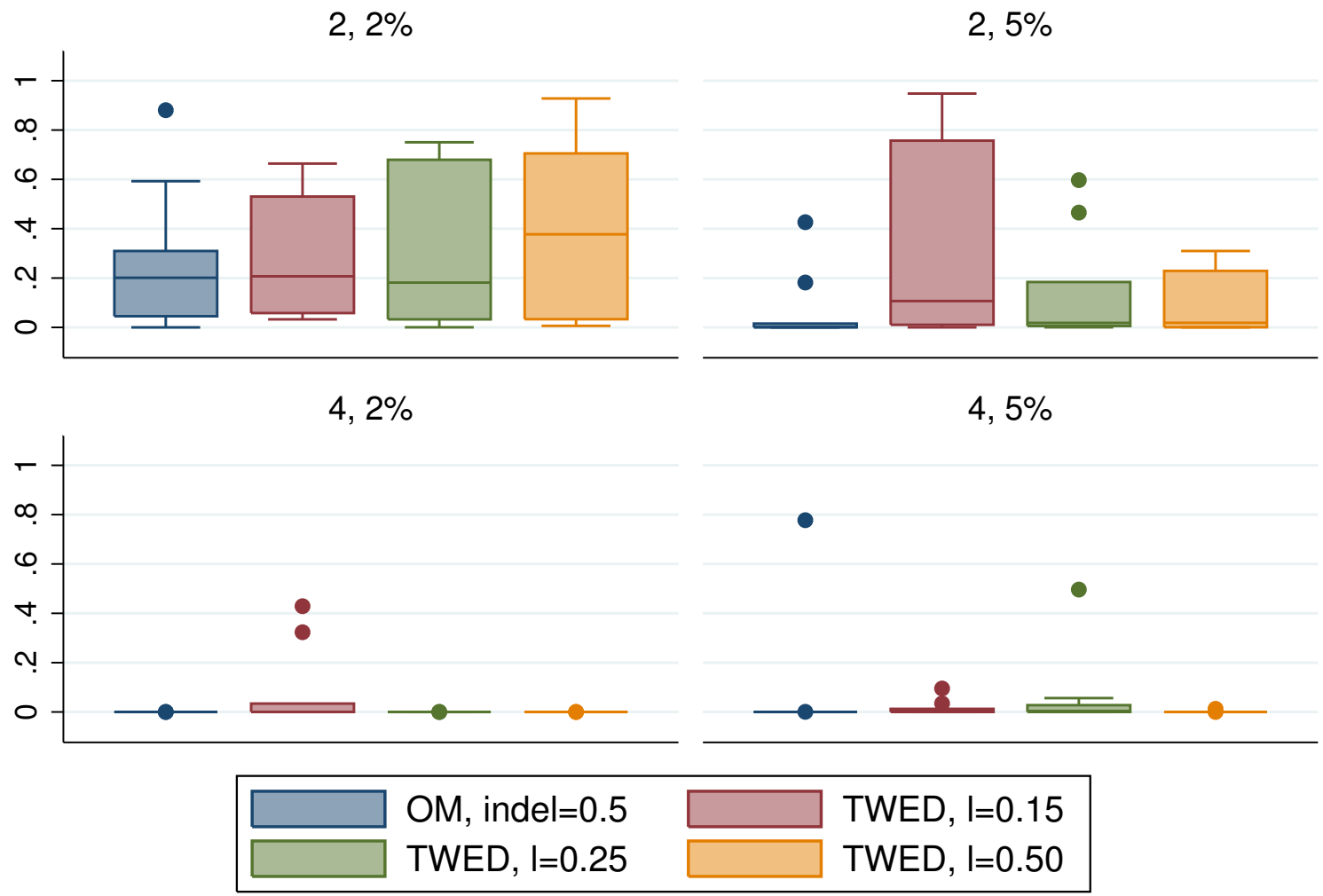
Brendan Halpin
Department of Sociology
University of Limerick

TWED and two-speed scenario



800 seqs, 50 reps

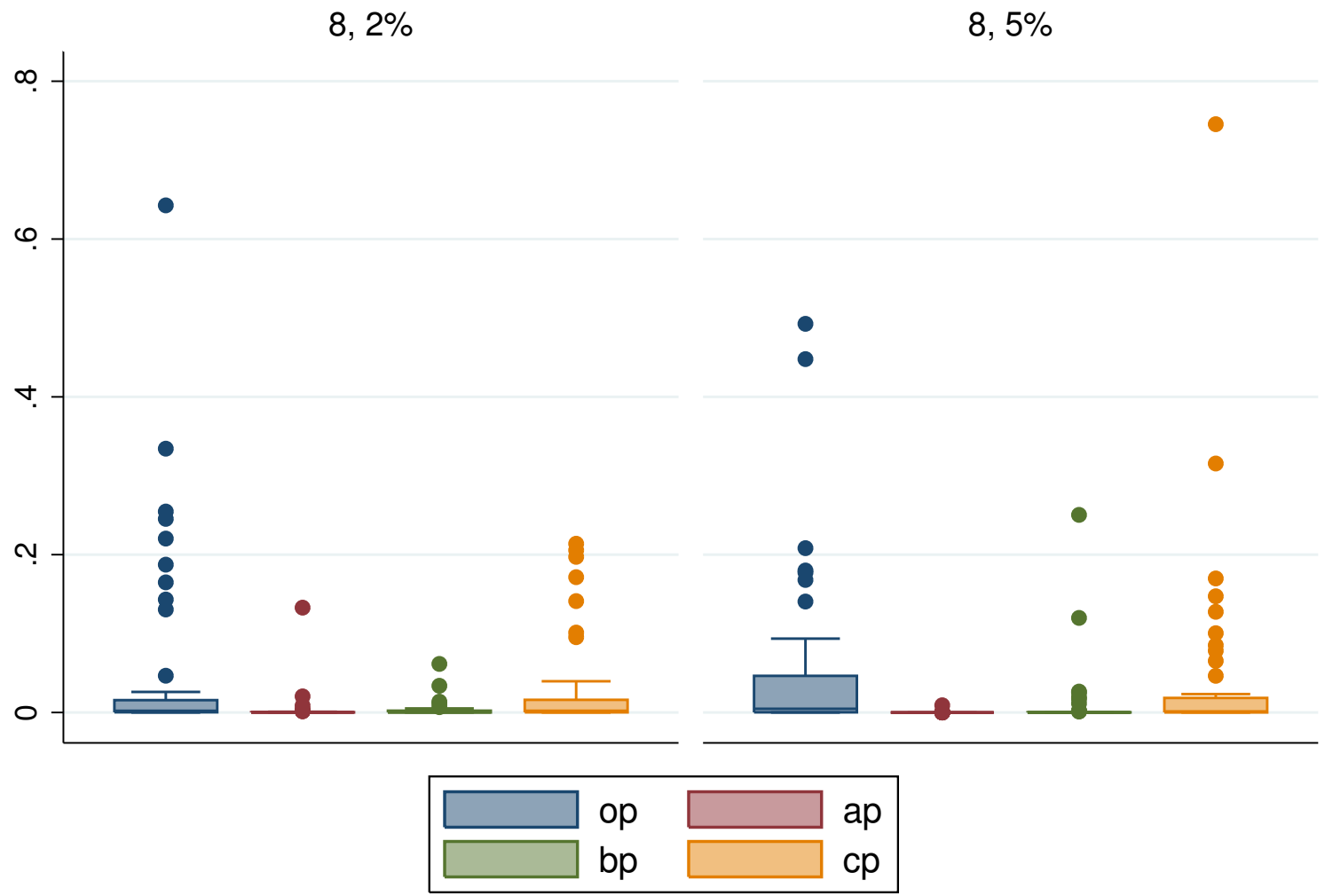
TWED and forced event scenario



800 seqs, 50 reps

Brendan Halpin
Department of Sociology
University of Limerick

TWED and fast-then-slow



Graphs by n and Threshold

Brendan Halpin
Department of Sociology
University of Limerick

TWED vs OM

- In the two-matrix scenario, TWED performs a little better than OM at low numbers of spells, but in contrast to OM improves as the number of spells increases.
- TWED is also competitive with OM in the two-speed and fast-then-slow scenarios
- The forced event scenario shows OM doing well: OM may be better at dealing with states at approximate times, TWED better at recognising sequence.
- Conclusion: The distance measure matters and OM is not the last word.

Brendan Halpin
Department of Sociology
University of Limerick

Taking transition rates seriously

- The foregoing simulations hinge on the structure of transition rates
- Let's apply this to real data:
  - First as a visualisation of the temporal structure of transitions
  - Second using this structure as a base for simulations against which we compare the reality.

Brendan Halpin
Department of Sociology
University of Limerick

Visualising the temporal structure of transition

- Sequences are generated by a complex, messy set of processes
- Individual difference, state dependence and temporal change are all likely to be important
- However, we can readily account for time by calculating the $M \times M \times (T-1)$ transition structure

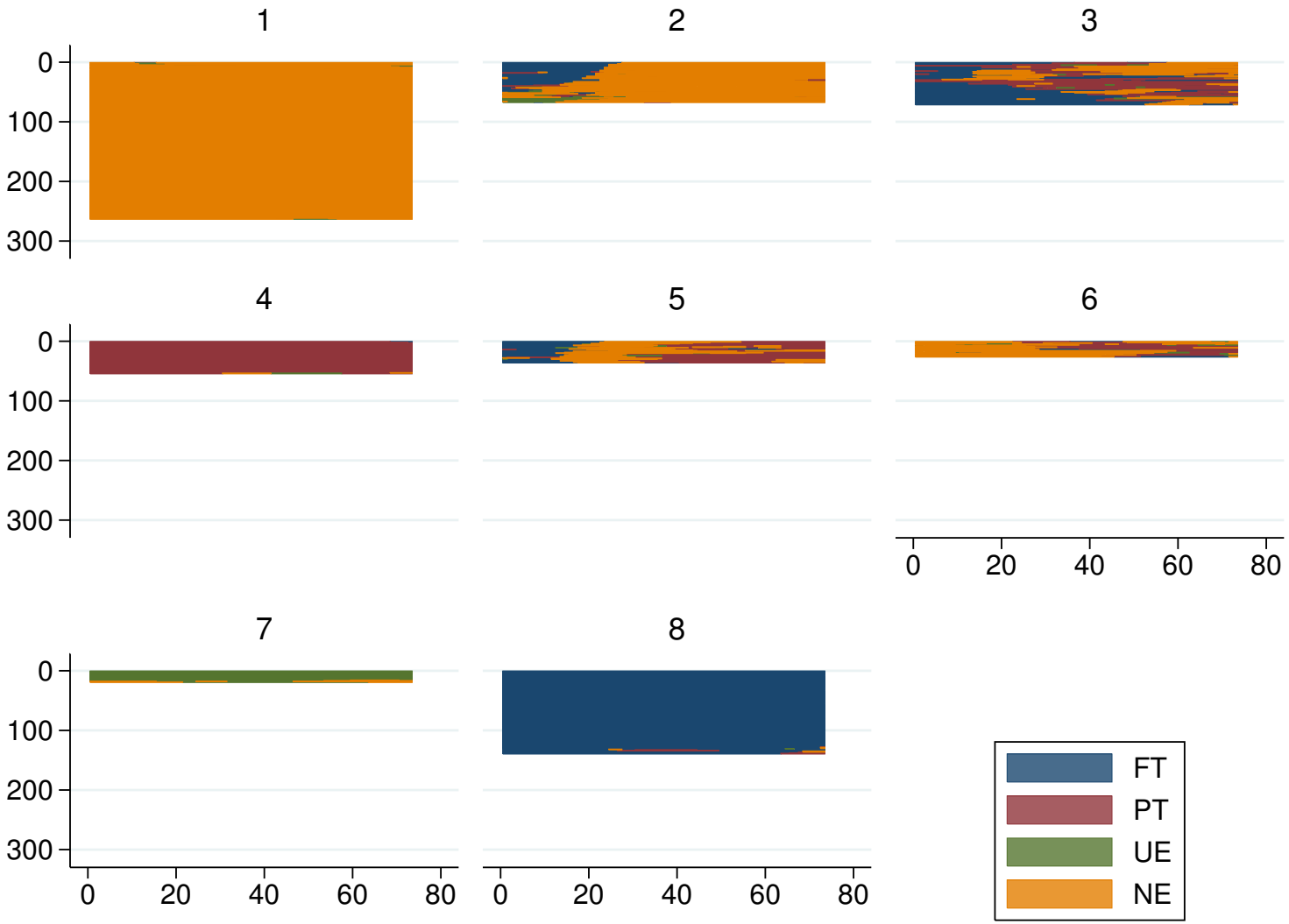
Brendan Halpin
Department of Sociology
University of Limerick

Example: Mothers' labour market sequence data

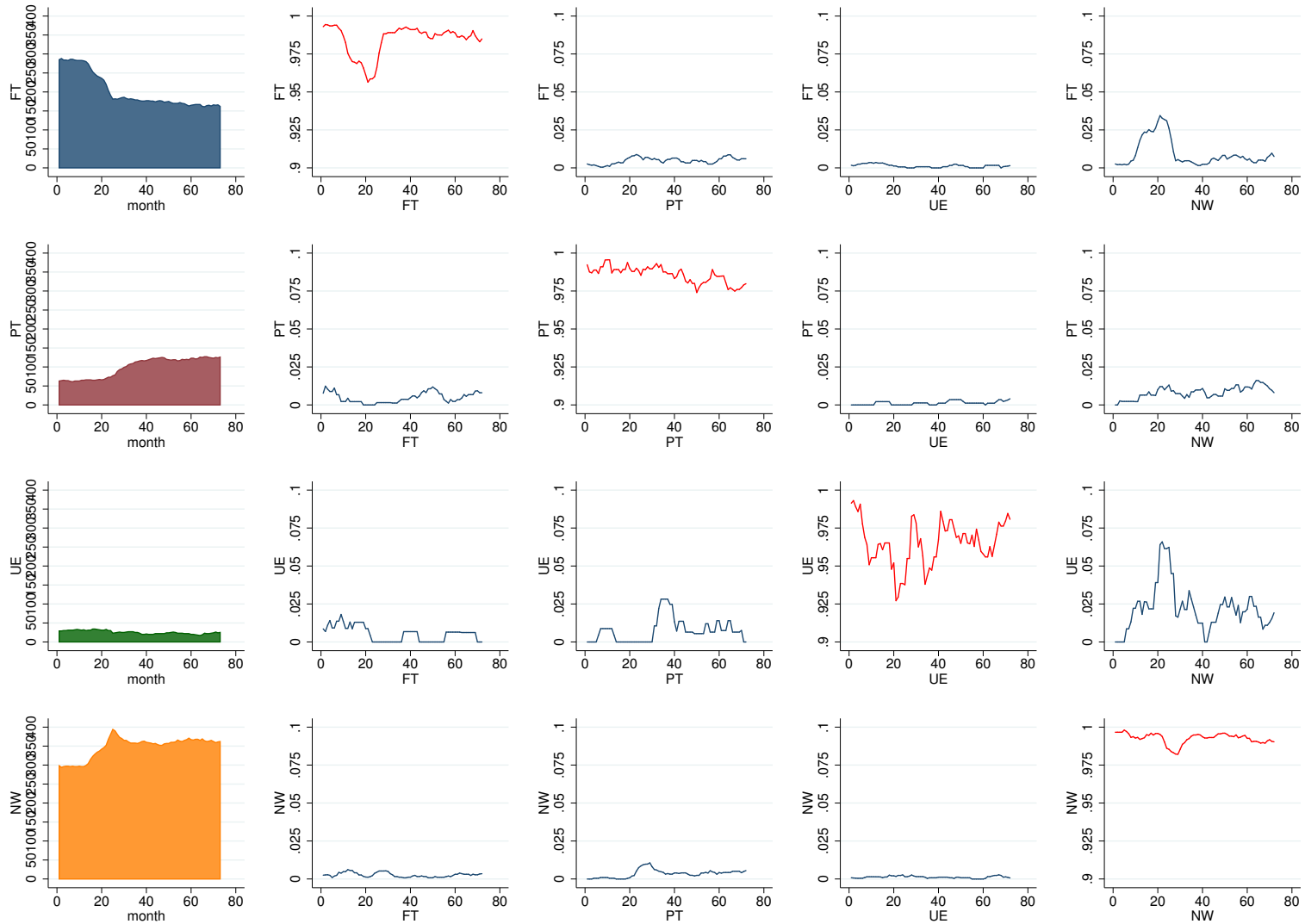
- Five years labour market history of women who have a birth at end of year 2
- Simple chronogram is informative but incomplete
- Cluster solution gives a digestible but messy overview

Brendan Halpin
Department of Sociology
University of Limerick

Mothers' labour market sequence 8-cluster solution

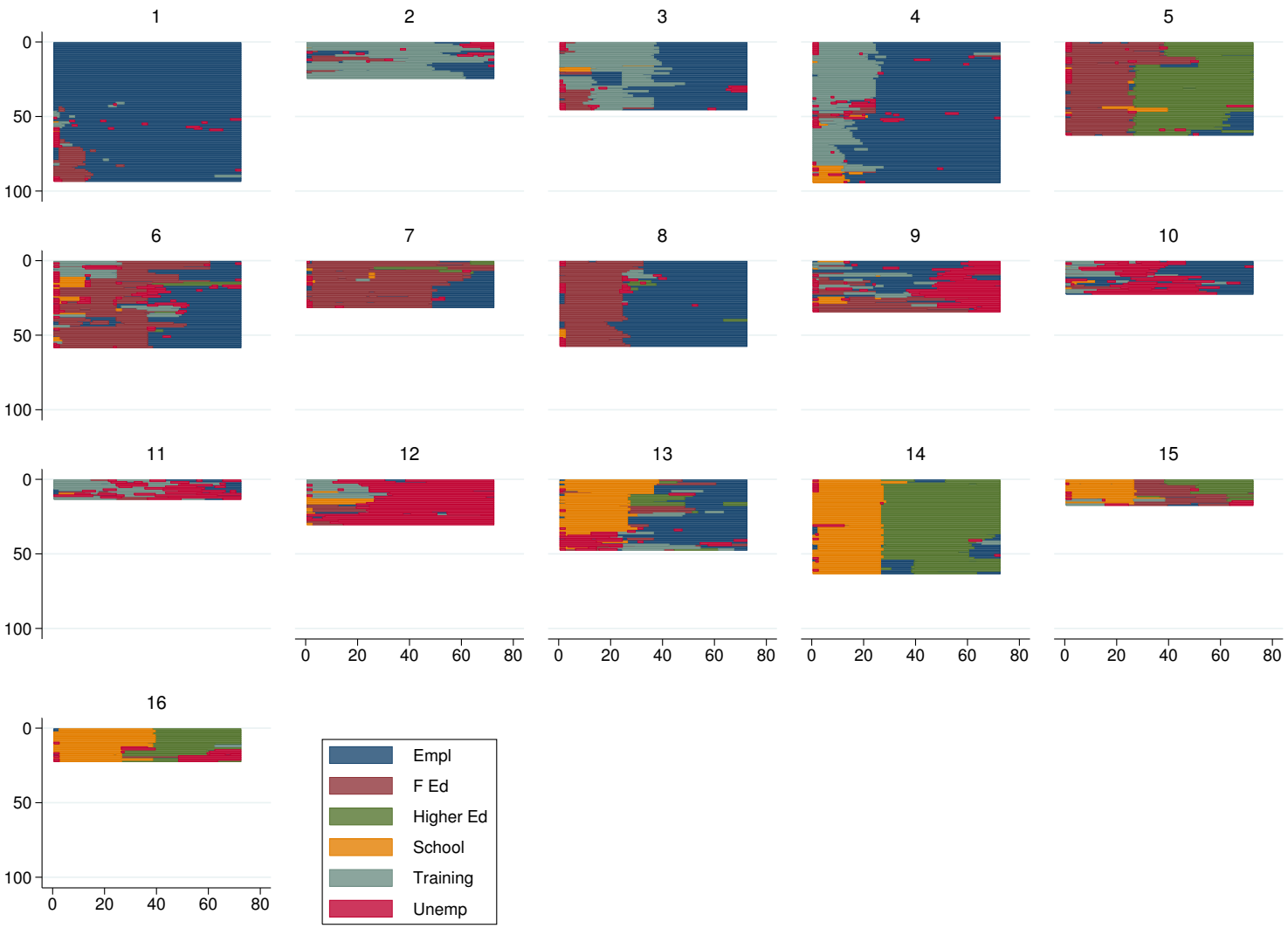


Temporal structure of transitions



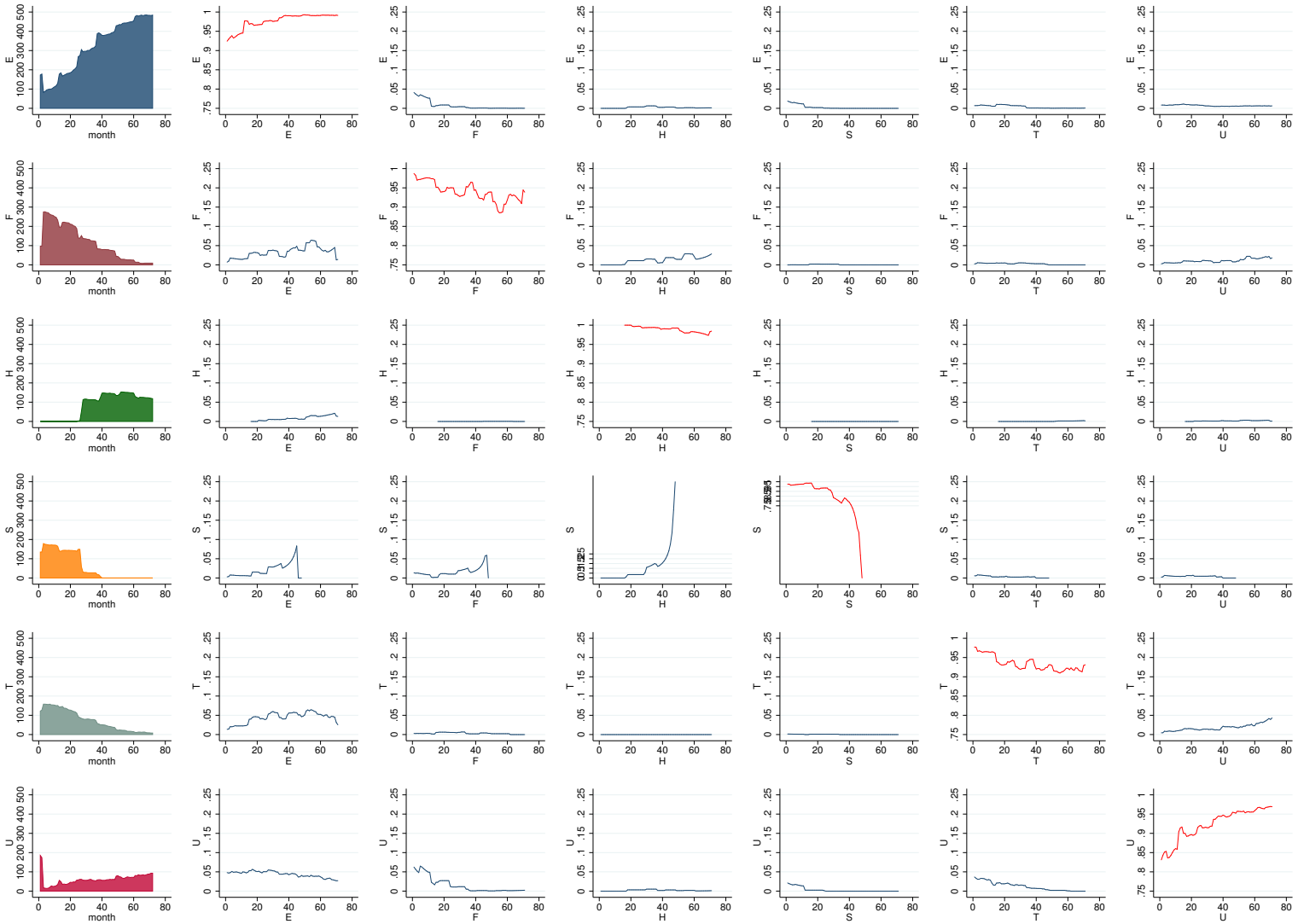
Brendan Halpin
Department of Sociology
University of Limerick

Cluster solution, MVAD data



Brendan Halpin
Department of Sociology
University of Limerick

Temporal structure, MVAD data



Brendan Halpin
Department of Sociology
University of Limerick

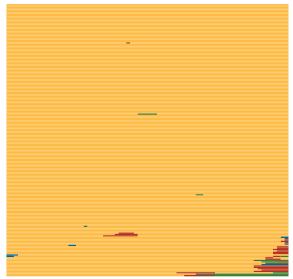
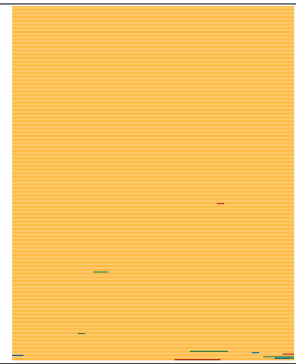
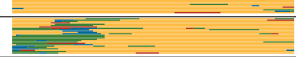
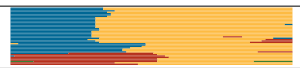
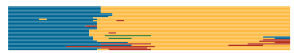
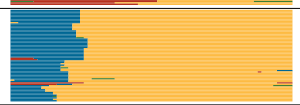
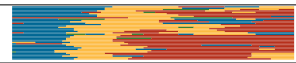
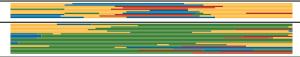
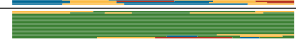
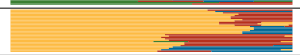
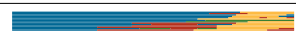
Comparing real and simulated, χ^2 test p-values

Groups		Births	MVAD	IMS	BHPS
2	mean	0.388	0.559	0.502	0.612
	median	0.384	0.617	0.509	0.650
4	mean	0.003	0.415	0.506	0.623
	median	0.000	0.377	0.502	0.694
8	mean	0.000	0.262	0.291	0.313
	median	0.000	0.188	0.153	0.150
16	mean	0.000	0.024	0.023	0.024
	median	0.000	0.000	0.001	0.002
32	mean	0.000	0.000	0.005	0.019
	median	0.000	0.000	0.000	0.002

Brendan Halpin
Department of Sociology
University of Limerick

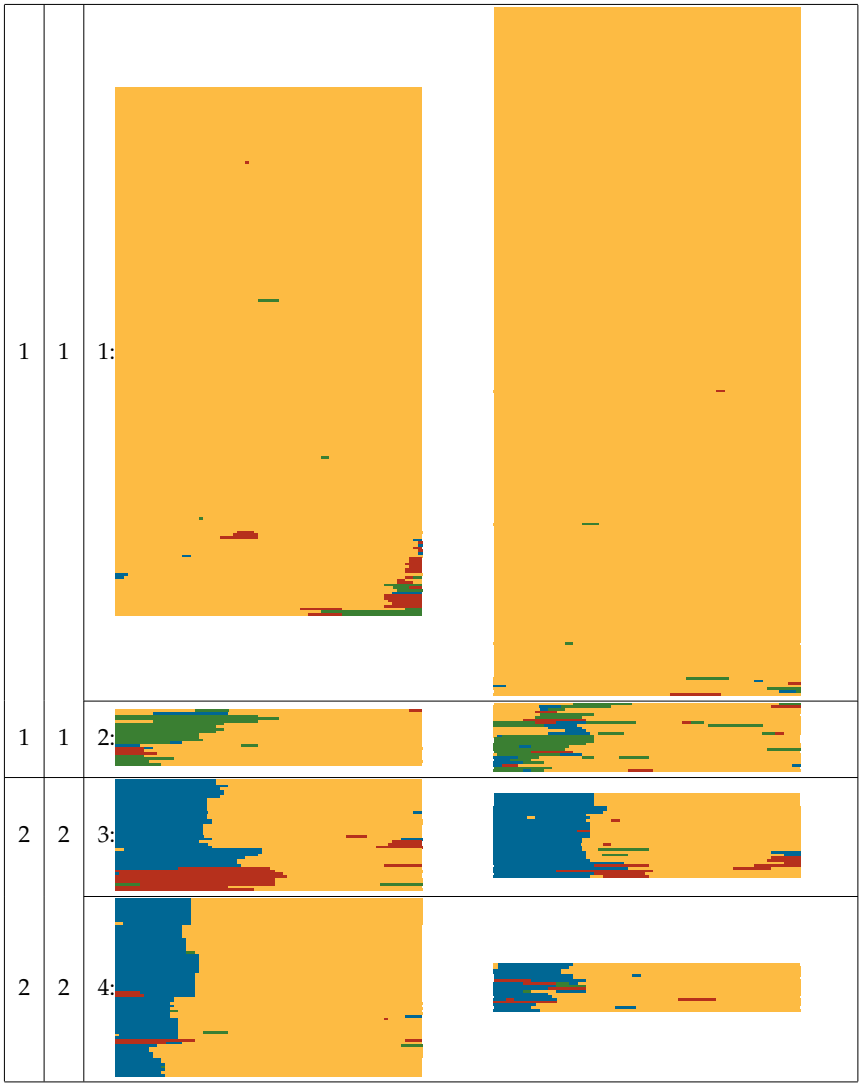
Simulated vs Real

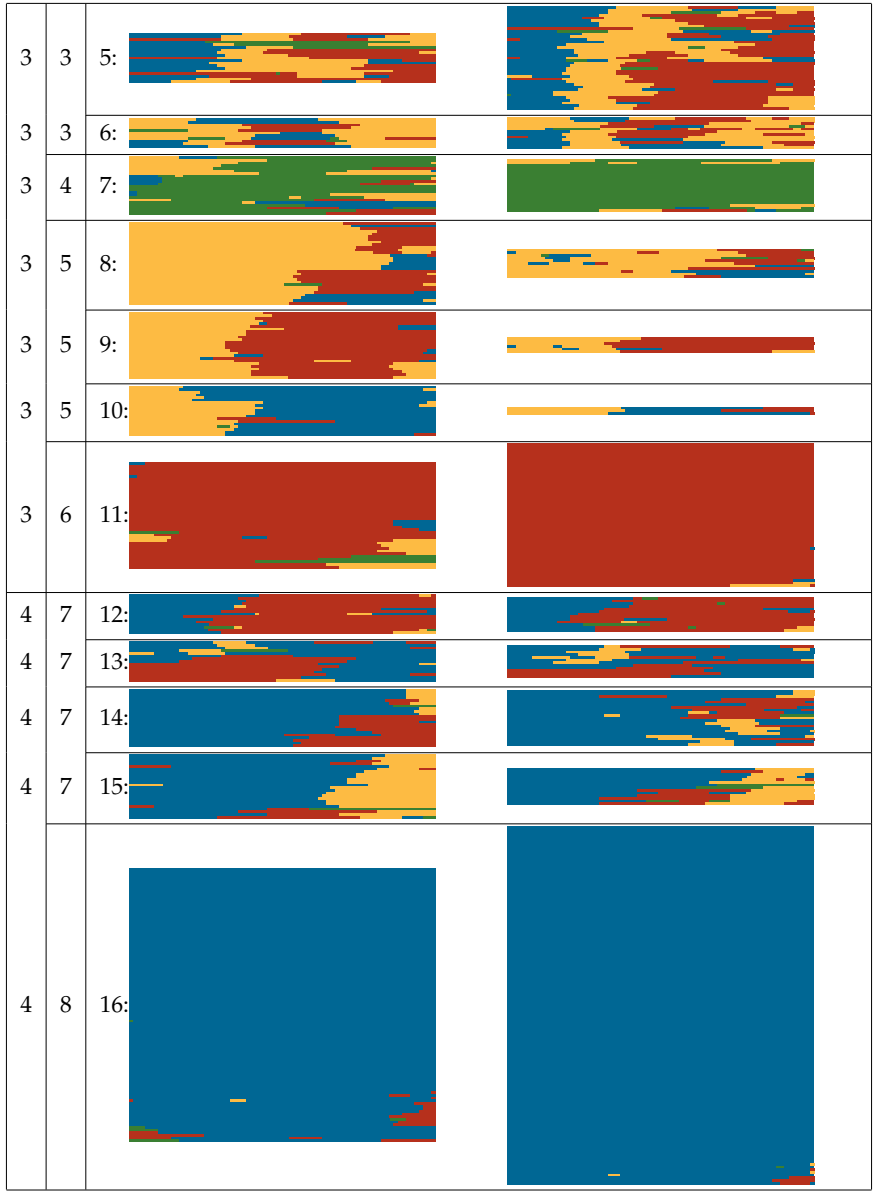
Single run of birth sequences

G2	G4	G8	G16	Simulated	Real
0.74	0.000	0.000	0.000 χ^2 p-values (this run)		
0.39	0.003	0.000	0.000 χ^2 p-values (mean over 50 runs)		
1	1	1	1:		
1	1	1	2:		
1	2	2	3:		
1	2	2	4:		
2	3	3	5:		
2	3	3	6:		
2	3	4	7:		
2	3	5	8:		
2	3	5	9:		
2	3	5	10:		
2	3	6	11:		
2	4	7	12:		
2	4	7	13:		
2	4	7	14:		
2	4	7	15:		
2	4	8	16:		

KEY: Full-time work Part-time work Unemployed Not available

Simulated vs Real





KEY: Full-time work Part-time work Unemployed Not available