SADI: Stata tools for Sequence Analysis

Brendan Halpin, University of Limerick

Stata User Group, London, 11/12 September 2014

http://teaching.sociology.ul.ie/seqanal/sadilondon.pdf

Outline

What is sequence analysis?

- A way of looking at time series as units
- Alternative to stochastic approaches that model the data generation process
- Origin in pattern recognition/machine learning; extensive application in molecular biology
- Advantage:
  - may capture structure that conventional approaches don't
  - provides a descriptive overview of complex data

How do we do sequence analysis?

- Calculate distance between pairs of sequences
- Use all pairwise distances to create empirical typologies
- Compare all sequences with a few ideal-typical sequences
- Compare pairs of sequences, e.g. spouses' time use; mothers' and daughters' fertility histories
- Address variability within groups (e.g., destandardisation of life course across cohorts)

How do we define distance?

- Count matching elements; identity at the same time
- Hamming distance: allow state space; full or partial similarity at the same time
- Aligning methods: full or partial similarity at the same or similar time
- Optimal Matching Algorithm uses token editing (substitution, insertion, deletion) to do such alignment
- OM evangelised extensively in sociology by Andrew Abbott

Controversy and alternatives

- How to determine substitution costs
- Do token sequences represent life course data well? (Hollister, 2009; Halpin, 2010; but see Halpin, 2014b)
- Some alternatives:
  - Dynamic Hamming (Lesnard)
  - Elzinga's combinatorial approaches
  - Time-Warp Edit Distance
  - For more detail see Halpin (2014b)

SADI: Sequence Analysis DIstance measures

- For a long time, little software for SA
  - Abbott's custom programme
  - Bioinformatics software for molecular sequence analysis
- Since then, a lot of options
  - Rohwer's TDA incorporated OM in mid/late 1990s
  - Kohler/Brzinsky-Fay/Luniak SQ for Stata since 2006
  - R Library Traminer since 2008
- SADI (first distributed 2007) takes a different approach to SQ

SADI compared to SQ

- Uses C plugins
  - Good: c 50X faster
  - Bad: problems of platform dependency, crashes
- Deals with duplicate sequences differently: consequences for cluster analysis
- More distance measures as well as OM
  - Hamming
  - Dynamic Hamming
  - Time Warp Edit Distance
  - Some of Elzinga's combinatorial measures

Key SADI components

- Distance measures
  - `oma`: Optimal matching distance
  - `hamming`: Hamming distance
  - `dynham`: Dynamic Hamming distance
  - `combinadd`: Elzinga's duration-weighted spell measure
  - `twed`: Time-warp edit distance
- Representing sequences
  - `stripe`: string representations of sequences
  - `chronogram`: state distribution over time
  - `trprgr`: time-series of transition rates
- Comparing solutions
  - `corrsqm`: correlation between pairwise distance matrices
  - `permtab`: tabulate cluster solutions
  - `ari`: Adjusted Rand Index, assess agreement of cluster solutions

A worked example: mothers' labour market histories

- Data derived from BHPS work-life histories
- 6 years, mothers who have a birth at end of year 2
- Full and part-time employed, unemployed, non-employed
- Unusual in that time keyed by event in middle, not start

Data structure

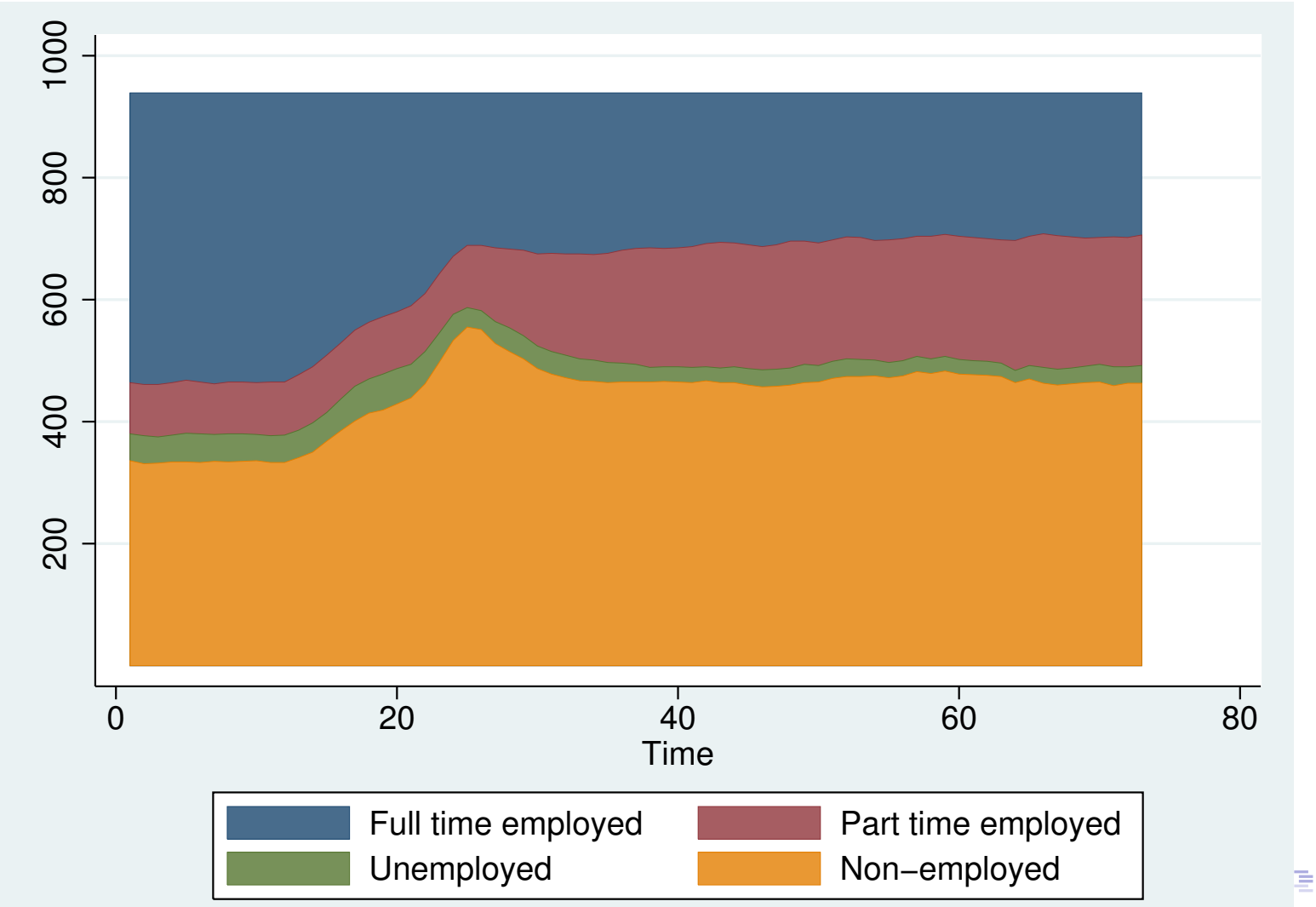
- Wide, one variable per month

```
. stripe state1-state72, gen(stripe) symbols("FPun")
. list stripe in 1/10, clean

                                                                                  stripe
  1.   nnnnnnnnnnnFFuuuunnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
  2.   FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFnnn
  3.   FFFFFFFFFFFFFFFFFFFFFFnnnnuuunnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
  4.   FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFPPPPPPPPPPPPPPPPPPPPPPPPPPPnnnnnnnnnnnnnn
  5.   FFFFFFFFFFFFFFFFFFFFFFnnnnnnnnnnFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFnnnn
  6.   FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFnnnnnnnnnnnnnFFFFFFFFFFFF
  7.   FFFFFFFFFFFFFFFFFFFFFFFnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnPPPPPPPPPPPPPPPPPP
  8.   FFFFFFFFFFFFFFFFFnnnnnnnnnnnFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
  9.   nnnnnnnnnnnnnnnnnnnnnnPPPPPPPPPPPPPFFFFFFFFFFFFnnnnnnnnnnnnnnnnnnnnuuuun
 10.   FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFnnnFFFFFFFFFFFFFFFFFFnnnnnnnnnnn
```

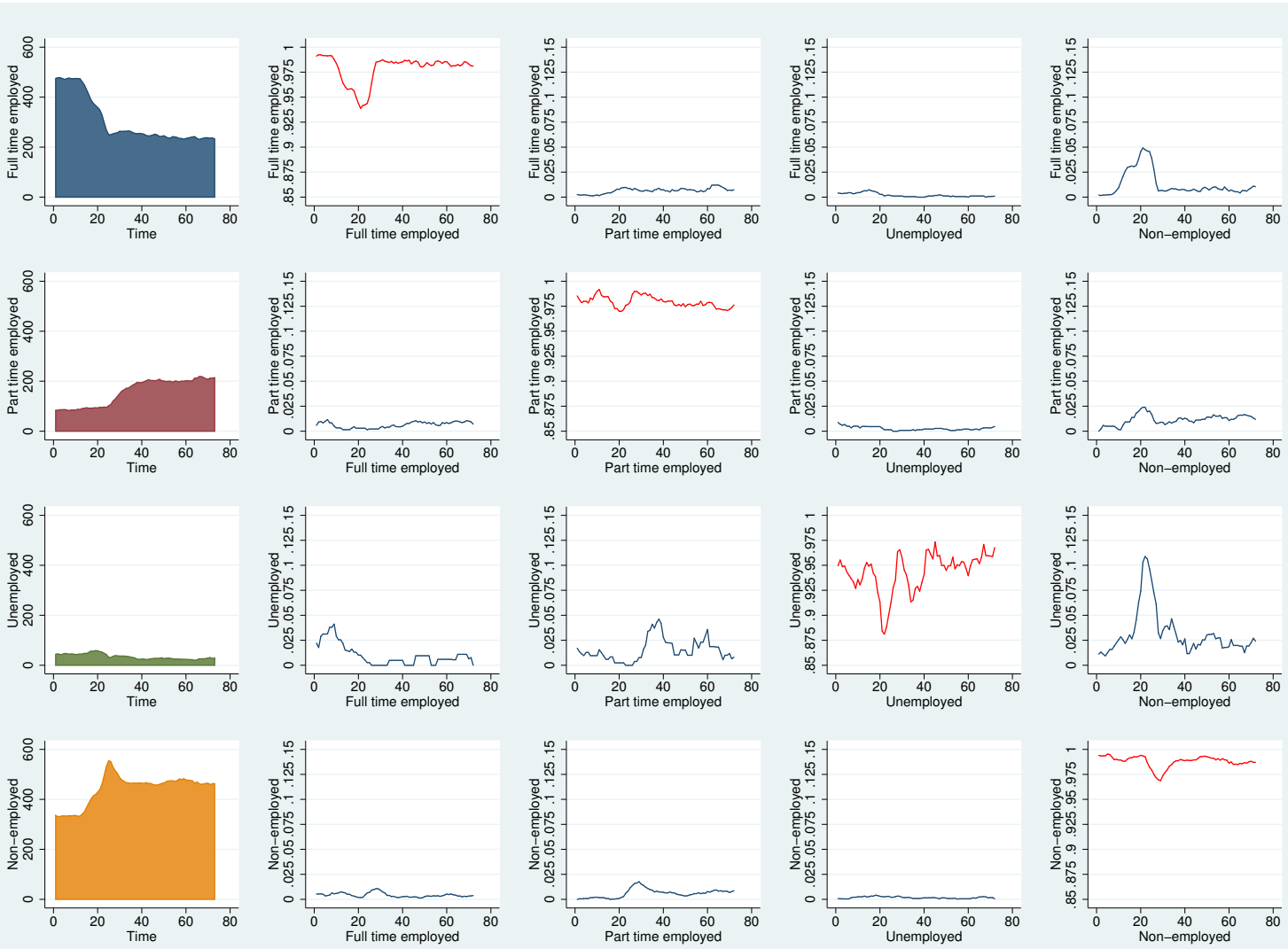
Chronogram: state distribution summary

```
. chronogram state*, id(pid)
```



trprgr: transition rate time-series

```
. trprgr state*, id(pid) gmax(575) floor(0.85) ceiling(0.15)
```

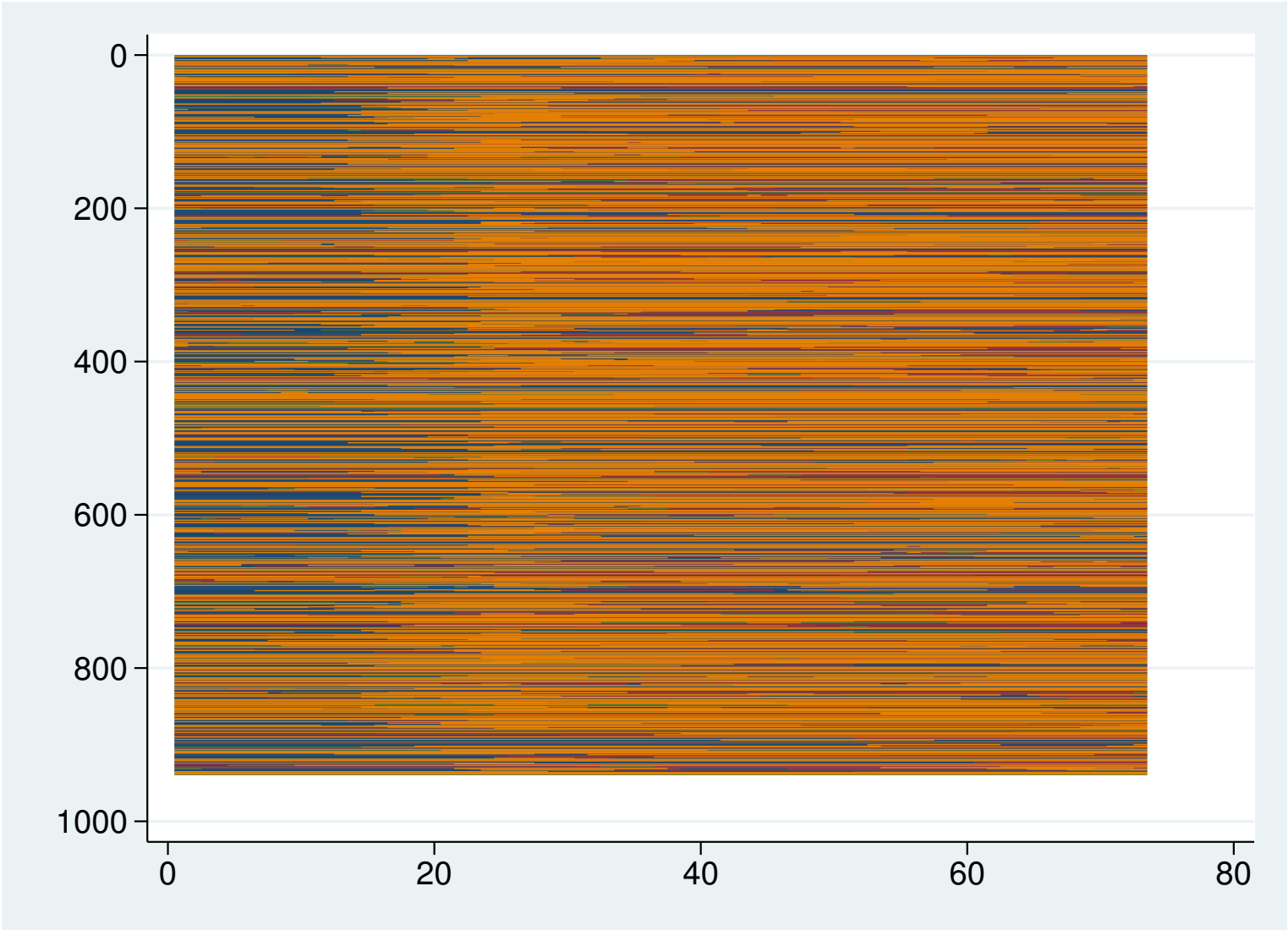


Indexplot, (SQ)

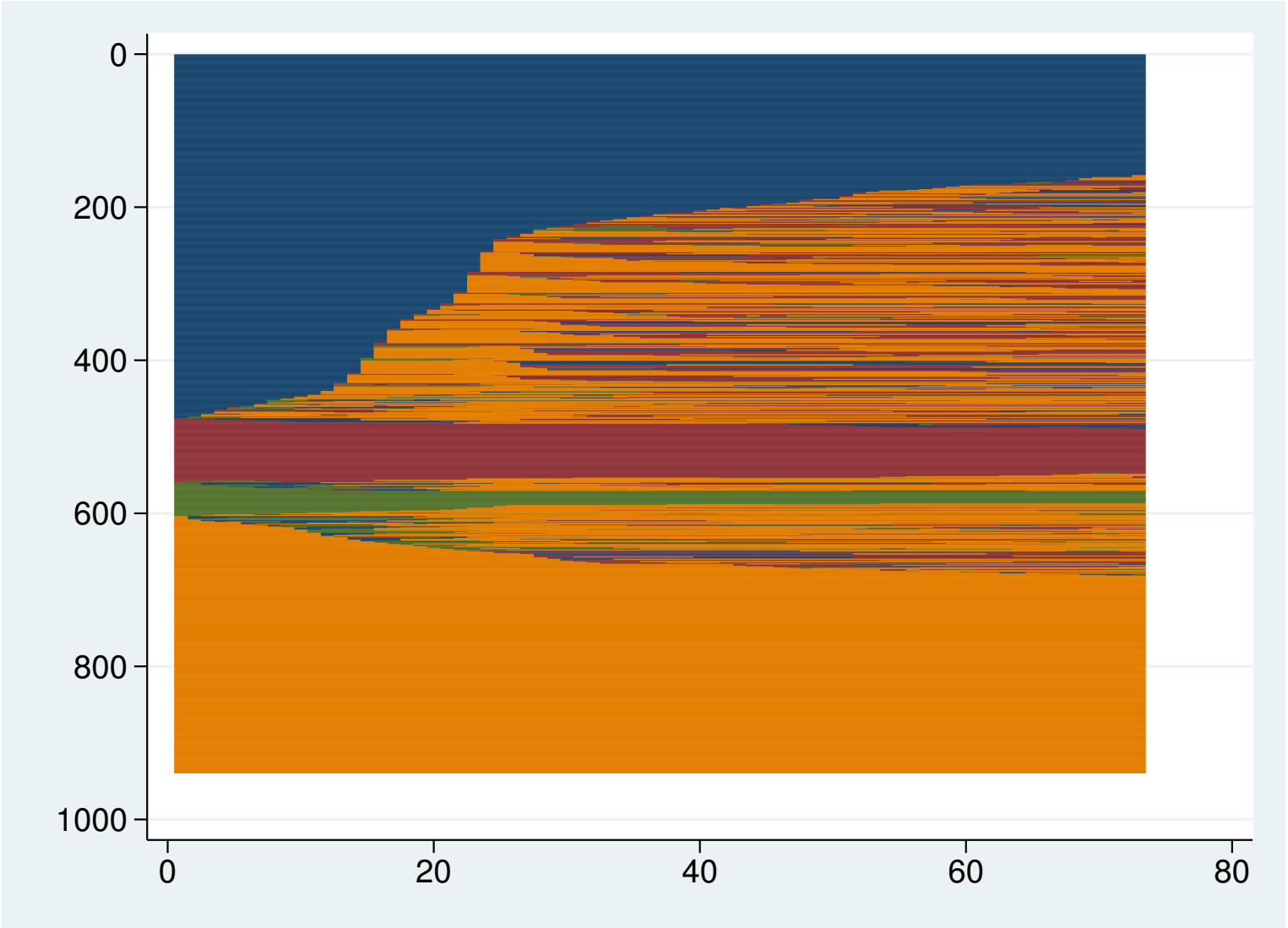
```
. reshape long state, i(pid) j(t)
. sqset state pid t
. sqindexplot, legend(off) overplot(100)
```

- This will generate a plot in "lexical" order
- Next graph is in random order, for a comparison

Indexplot, without order



Indexplot, lexically ordered



Optimal matching

- Let's define a simple state space: `F---P---u---n`
- This is represented as a substitution matrix:

```
. matrix sm = (0,1,2,3 \ ///
               1,0,1,2 \ ///
               2,1,0,1 \ ///
               3,2,1,0)
. oma state1-state72, subs(sm) indel(1.5) pwd(oml) len(72)
Normalising distances with respect to length
(0 observations deleted)
415 unique observations
```

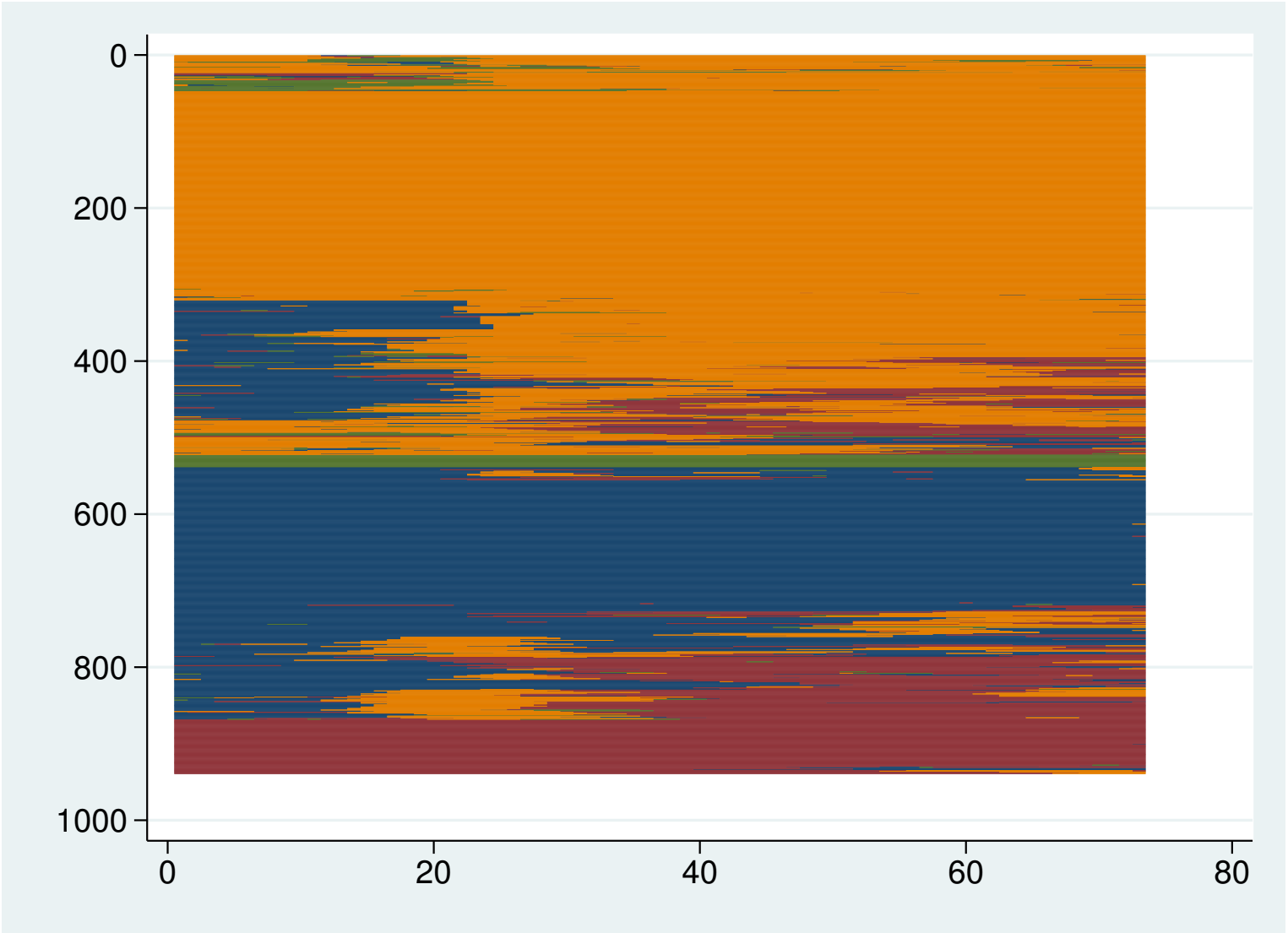
- *indel* cost 1.5 is half max substitution cost, as low as possible

Clustering the pairwise distances

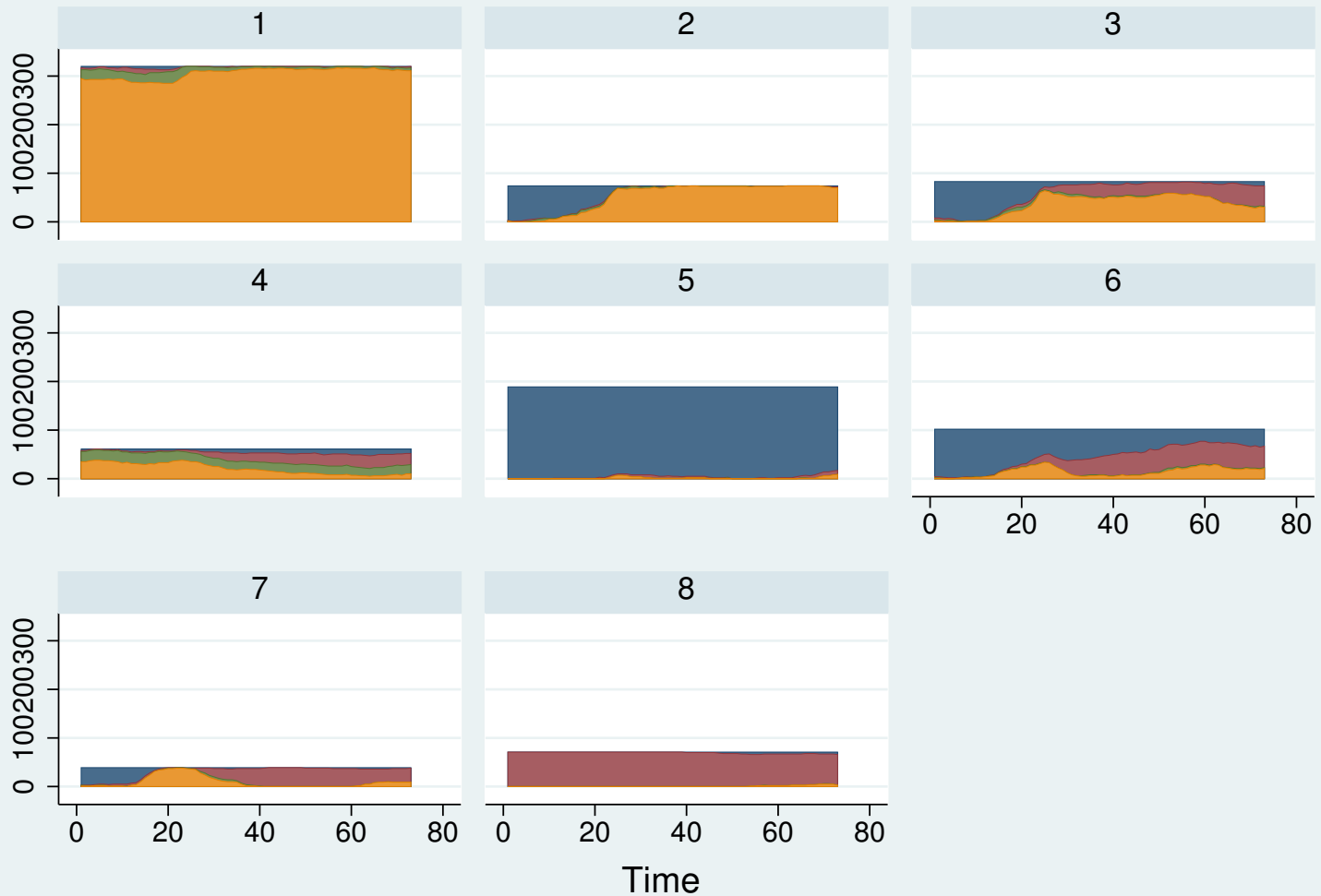
```
. clustermat wards oml, add
. cluster generate ol = groups(8 999), ties(fewer)
. tab ol8
```

ol8	Freq.	Percent	Cum.
1	320	34.08	34.08
2	74	7.88	41.96
3	83	8.84	50.80
4	61	6.50	57.29
5	189	20.13	77.42
6	102	10.86	88.29
7	39	4.15	92.44
8	71	7.56	100.00
Total	939	100.00	

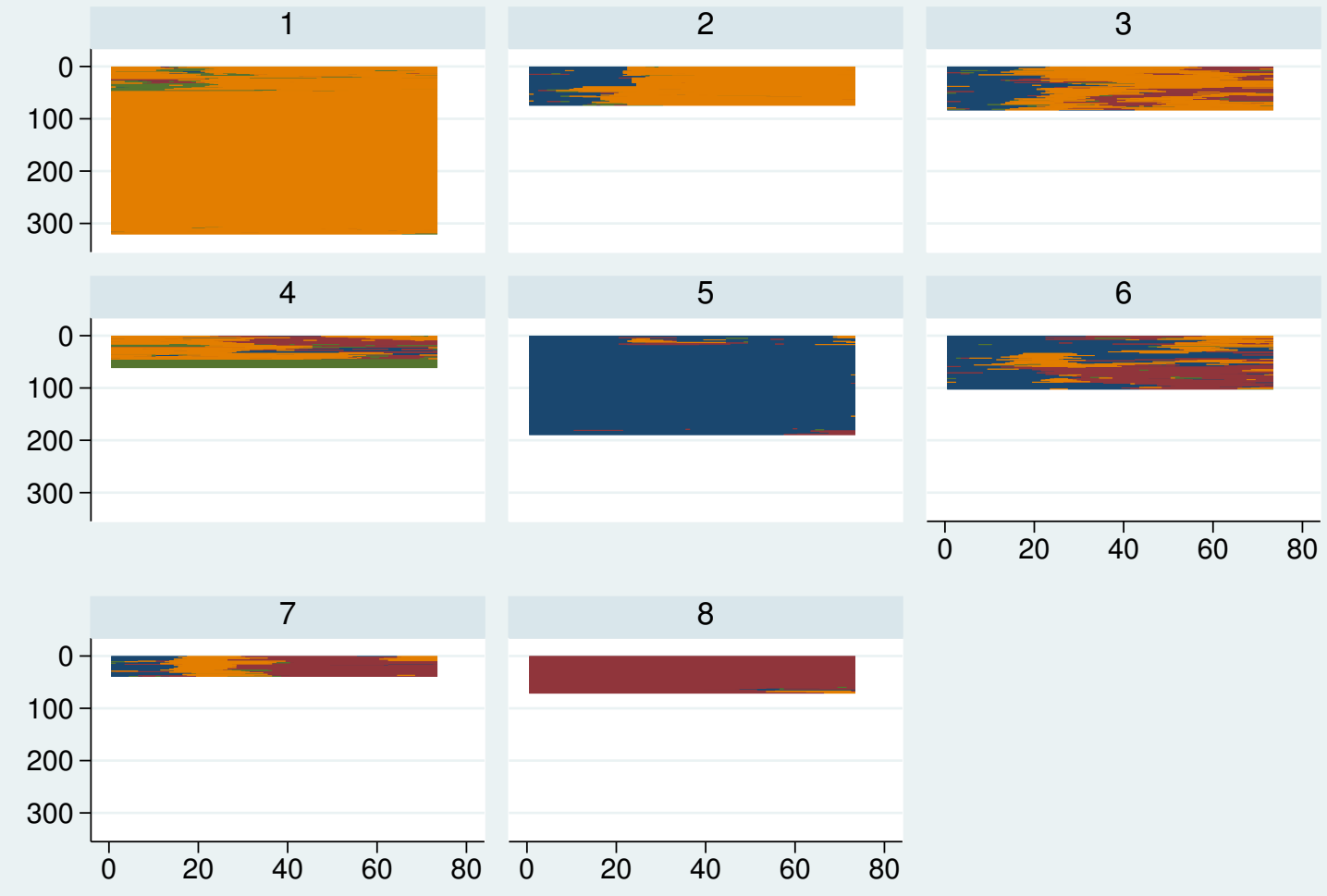
Indexplot, dendrogram order



Chronogram by cluster



sqindexplot by cluster with dendrogram order



Parameterisation

- Setting substitution and *indel* costs is difficult
- No theory, somewhat controversial
- I like to understand it as mapping a perspective on state-space onto trajectory-space
- However, changing the parameters changes the results

Two contrasting cost setups

```
. matrix sm = (0,1,2,3 \ ///
               1,0,1,2 \ ///
               2,1,0,1 \ ///
               3,2,1,0)
. matrix fl = (0,1,1,1 \ ///
               1,0,1,1 \ ///
               1,1,0,1 \ ///
               1,1,1,0)
. oma state1-state72, subs(sm) indel(1.5) pwd(oml) len(72)
. oma state1-state72, subs(fl) indel(0.5) pwd(omf) len(72)
```

Permuting linear and flat solutions

- Command: `permtab ol8 of8`

Kappa max: 0.7742

Permutation vector

	1
1	1
2	2
3	3
4	7
5	4
6	5
7	6
8	8

Permuted table:

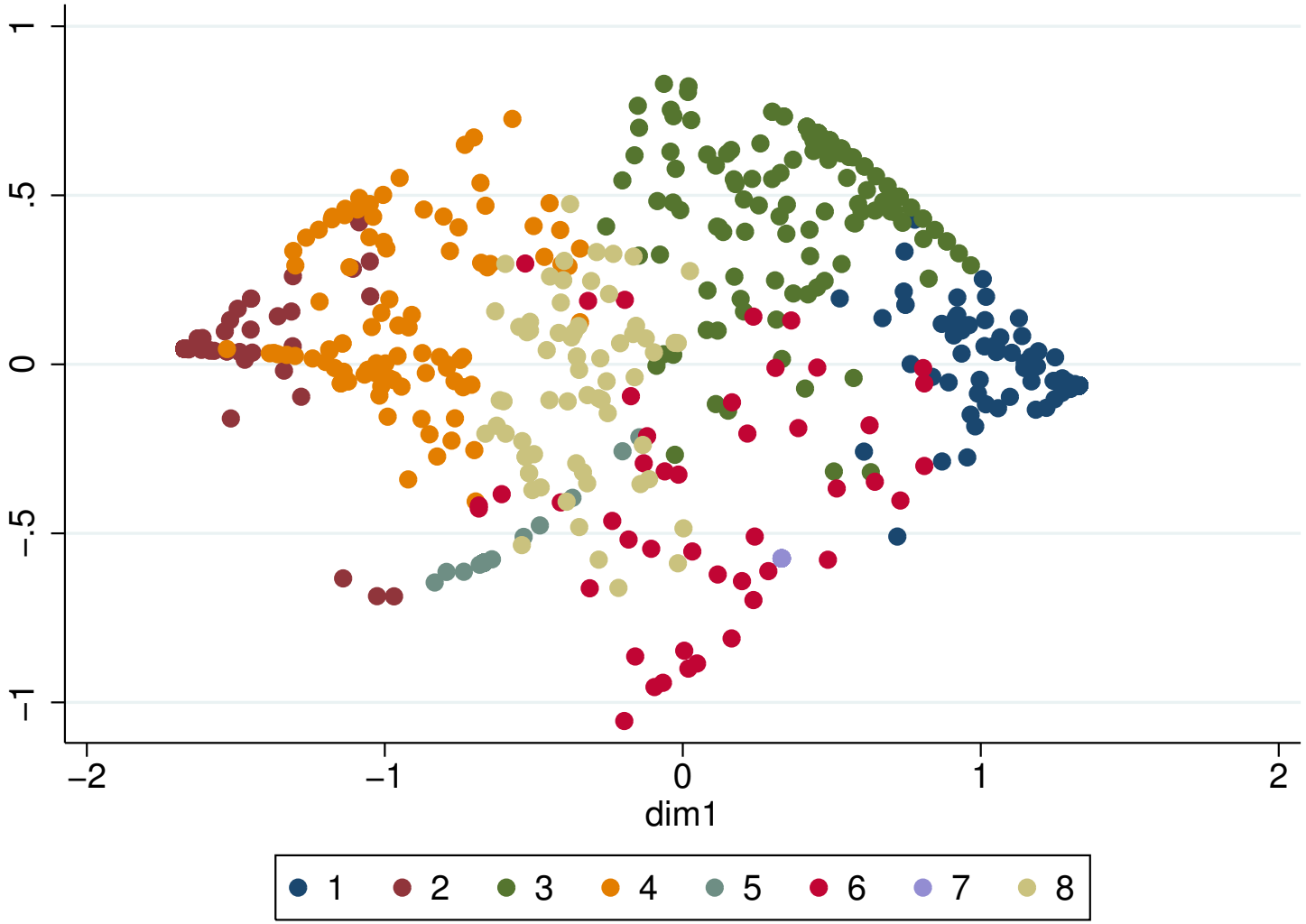
	1	2	3	4	5	6	7	8
1	293	26	1	0	0	0	0	0
2	1	72	1	0	0	0	0	0
3	0	3	76	0	0	2	2	0
4	5	0	0	16	0	14	24	2
5	0	0	0	0	180	9	0	0
6	0	0	10	0	39	21	32	0
7	0	0	0	0	0	1	38	0
8	0	0	0	0	0	0	0	71

Correlations of distances

- Summary based on `corrsqm` *mat1 mat2*, `nodiag`

Hamming linear	1.000	0.855	0.995	0.850	0.860	0.855	0.045
Hamming flat	0.855	1.000	0.850	0.987	0.998	1.000	0.094
OM linear	0.995	0.850	1.000	0.859	0.852	0.850	0.031
OM flat	0.850	0.987	0.859	1.000	0.980	0.987	0.066
TWED linear	0.860	0.998	0.852	0.980	1.000	0.998	0.127
TWED flat	0.855	1.000	0.850	0.987	0.998	1.000	0.093
X/t	0.045	0.094	0.031	0.066	0.127	0.093	1.000

Is clustering robust? Check with MDS



Discrepancy

- Studer et al's "discrepancy" measure gives us an alternative to cluster analysis
- Analogy to ANOVA and R-squared
  - TSS is the distance to the centre of gravity of the whole matrix
  - RSS is the distance to the centre of gravity of the partition
- Simple way to test for association between distance and a categorical variable

By Date of Birth, OM and X/t

```
. discrepancy dob, distmat(oml) id(pid) niter(5000)

Discrepancy based R2 and F, 5000 permutations for p-value

             | pseudo R2   pseudo F    p-value
-------------+-------------------------------
         dob |  .1439802   52.42148      .0002

. discrepancy dob, distmat(xts) id(pid) niter(5000)

Discrepancy based R2 and F, 5000 permutations for p-value

             | pseudo R2   pseudo F    p-value
-------------+-------------------------------
         dob |  .0693522   23.22551      .0658
```

Good and bad of plugins

- Statacorp encourages use of Mata over plugins
- But sometimes plugins are preferable
  - faster when doing loop-intensive calculations (x50)
  - access existing external code and libraries
  - implement algorithms and data structures not available (or slow) in Mata
    - e.g. recursive enumeration of subsequences
    - hashtable data structure
- Downsides
  - need to compile separately for numerous platforms
  - can crash Stata
  - C can be a nightmare!

Compiling for multiple platforms

- The main platforms for Stata seem to be:
  - Windows 64-bit
  - Windows 32-bit
  - MacOS (Intel CPU)
  - Linux 64-bit
  - Linux 32-bit
- From Linux64 it is possible to cross compile for Windows and Linux, 32 and 64 bit
- Cross compilation for Mac is difficult, but may be possible
- Compiling on Mac and on other Unix is straightforward

Cross-compilation on 64-bit Debian

- Load these packages (other distributions are analogous)

```
apt-get install mingw32
apt-get install mingw-w64
apt-get install libc6-dev-i386
```

- Then compile:

```
# Linux 32
gcc -m32 -fPIC -shared -DSYSTEM=OPUNIX stplugin.c example.c -o example.plugin
# Linux 64
gcc -m64 -fPIC -shared -DSYSTEM=OPUNIX stplugin.c example.c -o example.plugin
# Windows 32
i586-mingw32msvc-cc -shared -DSYSTEM=STWIN stplugin.c example.c -o example.plugin
# Windows 64
x86_64-w64-mingw32-gcc -shared -DSYSTEM=STWIN stplugin.c example.c -o example.plugin
```

MacOS

- On Mac, using gcc

```
gcc -bundle -DSYSTEM=APPLEMAC stplugin.c example.c -o example.plugin
```

(thanks to Glenn Hoetker, Arizona, for help compiling for Mac)

Installation

- For SADI

```
net from http://teaching.sociology.ul.ie/sadi
net install sadi
```

- SADI requires `moremata`

```
ssc install moremata
```

- For SQ, for indexplots

```
ssc install sq
```

Further reading

- Halpin, 2014a, *SADI: Sequence Analysis Tools for Stata*, WP2014-03, Dept of Sociology, University of Limerick, `http://www.ul.ie/sociology/pubs/wp2014-03.pdf`
- Halpin, 2014b, Three narratives of sequence analysis, in Bühlmann et al (eds), *Advances in Sequence Analysis*, Springer
- Halpin, 2012, *Sequence analysis of life-course data: a comparison of distance measures*, WP2012-02, Dept of Sociology, University of Limerick `http://www.ul.ie/sociology/pubs/wp2012-02.pdf`
- Studer et al., 2011, Discrepancy Analysis of State Sequences, *Sociological Methods and Research*, 40(3)
- Studer, 2012, *Étude des inégalités de genre en début de carrière académique*, Ch 2 "Comparaison des mesures de distance", `http://archive-ouverte.unige.ch/unige:22054`

This document

This document is available at
`http://teaching.sociology.ul.ie/seqanal/sadilondon.pdf`