Social Sequence Analysis

Brendan Halpin, Dept of Sociology, University of Limerick

Umeå, March 30 2017

Outline

- What is sequence analysis?
- Why it can be worth doing, and how it complements existing approaches
- Uses it has been put to
- Criticisms
- Future directions

Slides available at `http://teaching.sociology.ul.ie/umea`

Sequence Analysis

- What is sequence analysis?
  - Large and active research area
  - From Andrew Abbott in mid-late 1980s, to 2015 special edition of *Sociological Methodology*
- Focuses on linear data (such as lifecourse trajectories) as *sequences*, as wholes
- Usually proceeds by defining distances between pairs of sequences, creating empirical typologies, etc

A brief history of SA in Sociology

- Andrew Abbott's long evangelism
  - Abbott (1984) - earliest, argues for focusing on sequence as well as duration
  - Abbott and Forrest (1986) - Morris dancing
  - Abbott and Hrycak (1990) - careers of Baroque musicians
- Abbott's main point: focus on sequences as wholes as an alternative to "variable-based" sociology
- However, his main practical contribution was to introduce the OM algorithm to the social sciences

A brief history of SA in Sociology

- Andrew Abbott's long evangelism
  - Abbott (1984) - earliest, argues for focusing on sequence as well as duration
  - Abbott and Forrest (1986) - Morris dancing
  - Abbott and Hrycak (1990) - careers of Baroque musicians
- Abbott's main point: focus on sequences as wholes as an alternative to "variable-based" sociology
- However, his main practical contribution was to introduce the OM algorithm to the social sciences

James Coleman: 'No one's gonna pay any attention . . . as long as you write about dead German musicians' (Abbott, 2001, p. 13)

Some 1st wave adopters 1/2

- Stovel et al. (1996): A sequence-oriented analysis of career data from a British bank, showing a transition between a status-based and an achievement-based system, from 1890 to 1970.
- Wuerker (1996): Treats sequences of services interactions of mental health patients in Los Angeles. A small data set, but of interest because it uses a relatively uncommon form of trajectory.
- Halpin and Chan (1998): Analyses class careers of British and Irish men to age 35 using retrospective data.

Some 1st wave adopters 2/2

- Blair-Loy (1999): Women's careers in the finance industry; identifies change across cohort in opportunity and perspective.
- Han and Moen (1999): How life and work trajectories of couples are coordinated. Dyadic, not analyis of all pairwise distances: uses OM to generate a measure of intra-couple similarity.
- Stovel (2001): Not life-course: looks at county-level histories of lynching in the Southern US, drawing strongly on arguments from Abbott and others about the necessity of taking a sequence perspective on historical explanations.

2000 debate in SMR

- Position: Abbott and Tsay (2000)
- Critiques: Levine (2000) and Wu (2000)
  - is it sociologically meaningful?
  - how do we parameterise it?
  - does it have any advantages over conventional approaches?
- Response: Abbott (2000)

Key developments since

- Widespread in many fields, especially lifecourse related:
  - transition school to work, labour market, retirement, health outcomes, time use
  - Some focus on multiple domains, dyadic approaches, cohort change in average diversity
  - Much still uses clustering to develop empirical typologies
- See Aisenbrey and Fasang (2010) and Halpin (2013) for a summary
- Rather more activity in Europe than in US
- Two important conferences:
  - LaCOSA1 2012 on Sequence Analysis: Blanchard et al. (2014)
  - LaCOSA2 2016 on Sequence Analysis and related methods (Online proceedings: `https://lacosa.lives-nccr.ch/online-proceedings`)

Why do Sequence Analysis?

- Why would we want to do it
  - Holistic vs analytic?
  - Exploratory vs hypothesis testing?
  - Descriptive, visualisation
- Complexity of longitudinal processes hard to capture
- Complementary alternative to stochastic techniques which model data generation process

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 incomplete view
- SA clearly allows us visualise complex data; possibly allows us observe features that will otherwise be missed

Potentially complex 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
  - Clear exploratory advantages
  - possibility of detecting things that might not be detected otherwise

Timing, sequence, quantum

- Different things can be interesting
  - Timing: when things happen
  - Sequence: in what order do things happen
  - Quantum: how much time is spent in different states (Billari et al., 2006)

Non-holistic approaches

- Numerous non-holistic approaches exist
- Typically they will discard some aspect of the information in the data, and focus powerfully on another
- For instance, focus on
  - cumulated duration in states (how much but not when)
  - transition patterns between states (period-to-period but not overall)
  - time-to-event of leaving spell (spells, perhaps pooled, but lose sight of individual career).

Cumulative duration

- For instance, summarise trajectories in terms of cumulative time in each state
- Typically use as a predictor (e.g., proportion of time unemployed predicting later ill-health)
- Or as an outcome: variables measured earlier (e.g., school performance) predicting proportion of time unemployed.

Transition rate models

- Model rates of period-to-period change: e.g., monthly movement between labour market statuses
- Model origin–destination patterns: e.g., transition between class at entry to labour market, and class at age 35
- Markov models
- Very useful, good overview, can be descriptive or stochastic: tables make categorical data digestible
- Disadvantage: the focus on the t-1/t or t_0/t_T pattern means a loss of individual continuity
- Some potential to model longer Markov chains (Gabadinho, 2014)

Hazard-rate modelling

- Hazard-rate modelling is one of the dominant statistical alternative
- Either in terms of survival tables and curves (essentially descriptive)
- Or full stochastic models of the determinants of the hazard rate (Cox and/or parametric)
- Example: what characteristics speed up (or slow down) exit from unemployment?
- Very nice conceptual model of the temporal process
- Can test hypotheses
- Disadvantage: spell orientation, lack of whole-trajectory overview

Latent class analysis

- Latent class growth curve models
  - Where theory allows a developmental model of a quantitative outcome
  - Account for the structure of repeated measurement of individuals
  - Not so suitable for categorical variables
- Latent class models can be applied to careers
  - However, difficult to properly incorporate the longitudinality
  - Examples: Lovaglio and Mezzanzanica (2013); Barban and Billari (2012)

Hidden Markov Models

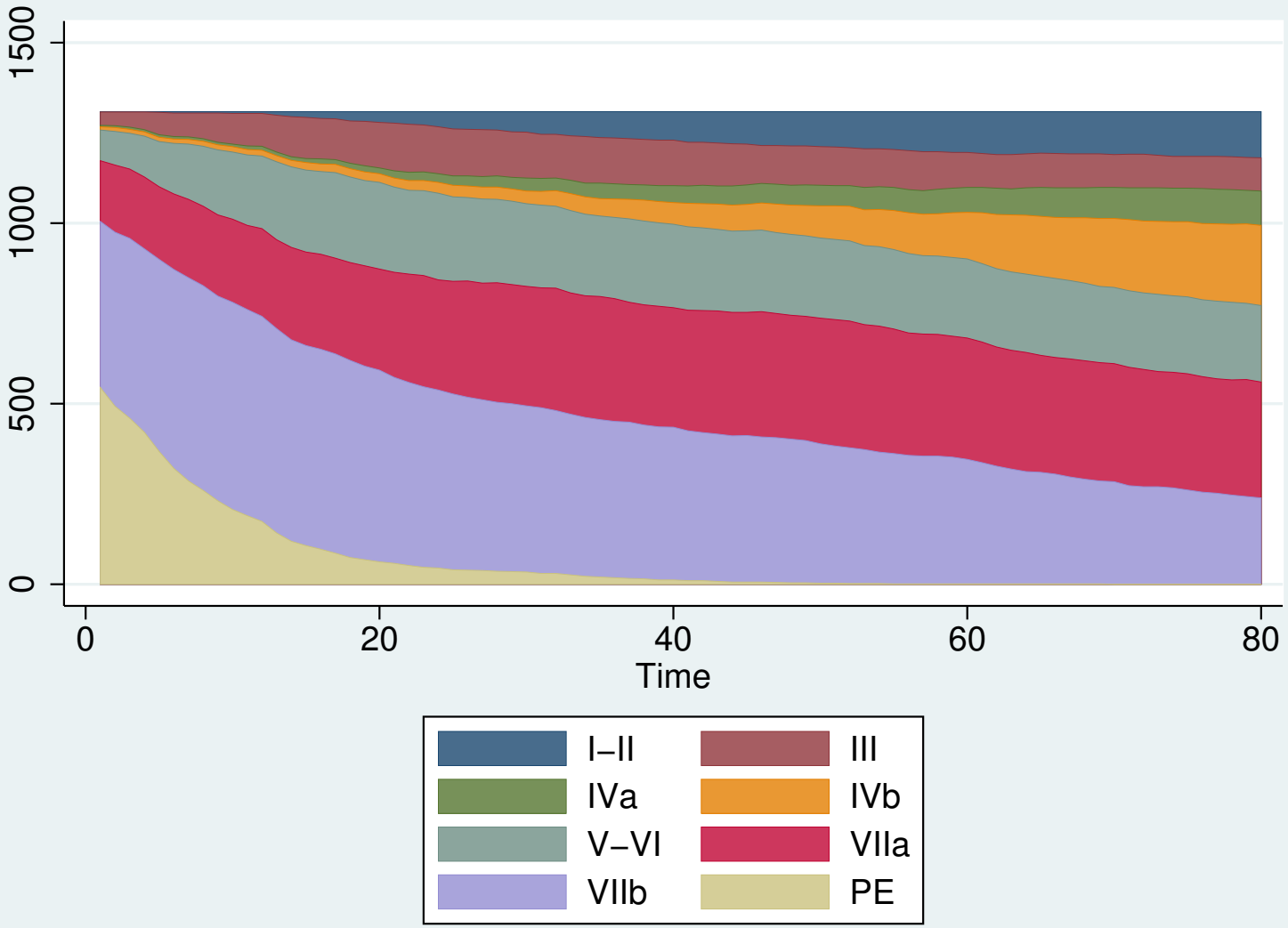
Helske et al at LaCOSA2 proposed a hidden Markov modelling approach to multi-channel sequence analysis

- unobserved states are probabilistically associated with observed states
- movement between the unobserved states can be modelled as a Markov process
- potentially a parsimonious & stochastic approach to modelling trajectories through a complex state space
- However, is computationally complex and can be unstable

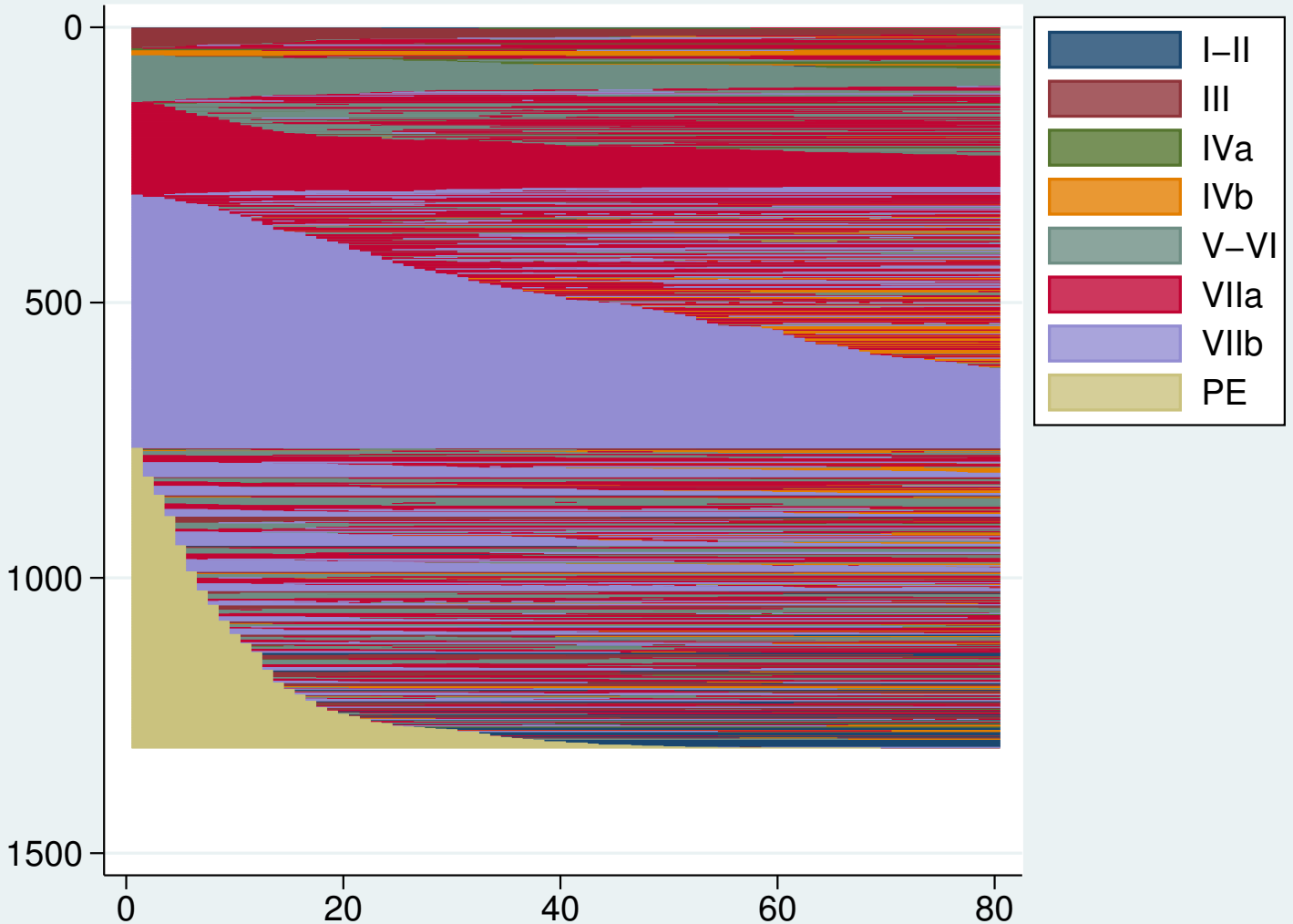
Worked example: 20th Century class careers in Ireland

- Data from the Irish Mobility Study, 1973
- Retrospect class careers of males, age 15-30, quarters
- 7 Goldthorpe class categories, plus a "pre-entry" state
- Primary goal: get an overview

Chronogram


0
500
1000
1500
0
20
40
60
80
Time
I–II
III
IVa
IVb
V–VI
VIIa
VIIb
PE

Raw indexplot



Parameterisation

- OM allows us to recognise similarity at the same time and (to a greater or lesser degree) displaced in time
- The analyst has to determine what consititutes similarity, and how easy to make displacement
- Similarity is expressed as a "substitution matrix", detailing symmetric distances between states
- Displacement is facilitated with a low "indel cost" (minimum is half maximum substitution cost)

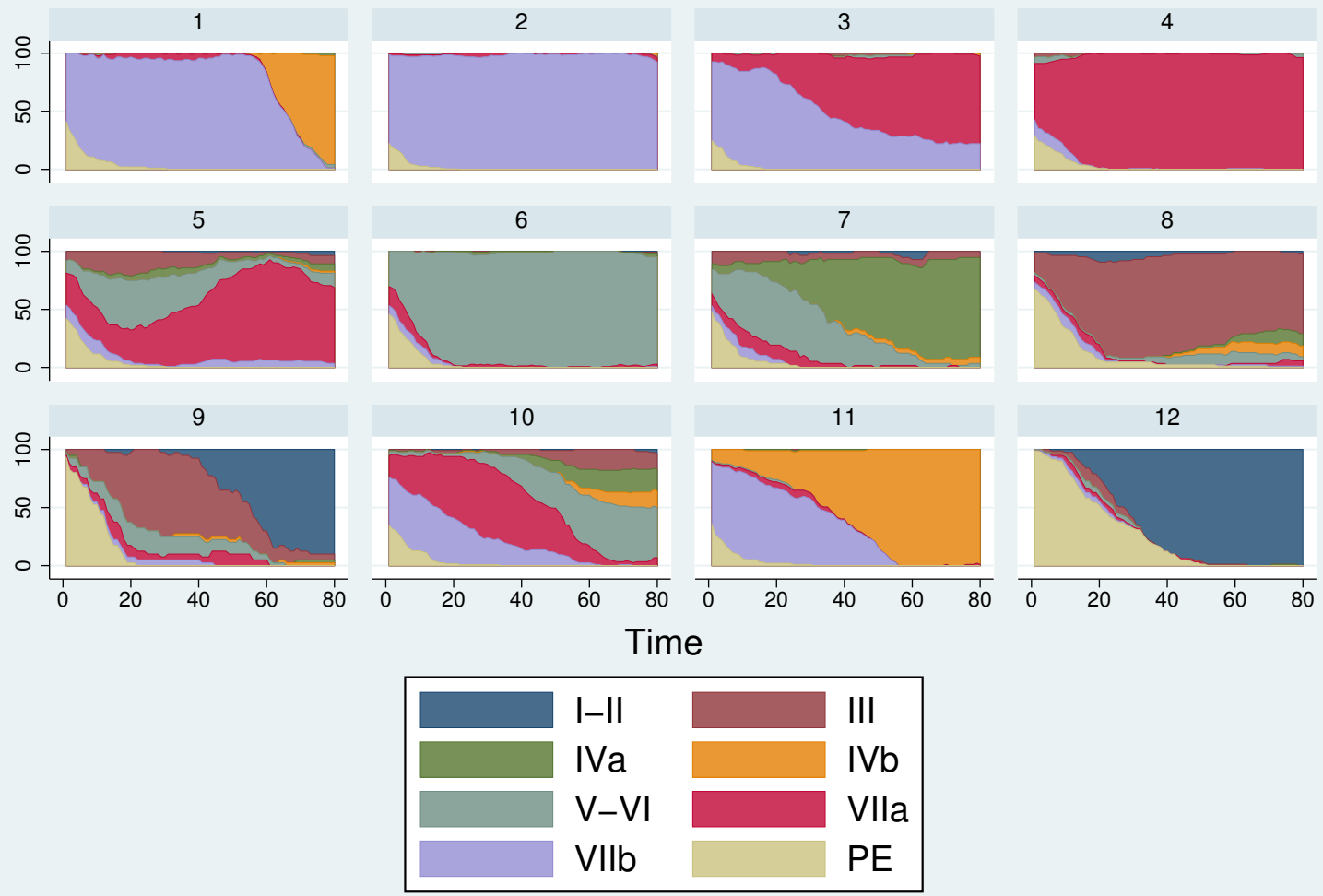
OM

- Running OM allows us to define distances between every pair of sequences
- We can use this to cluster sequences
- The indexplot ordered by the cluster analysis is much more informative
- The clusters may be useful for further analysis
  - What predicts membership of clusters?
  - Does cluster membership predict later outcomes?

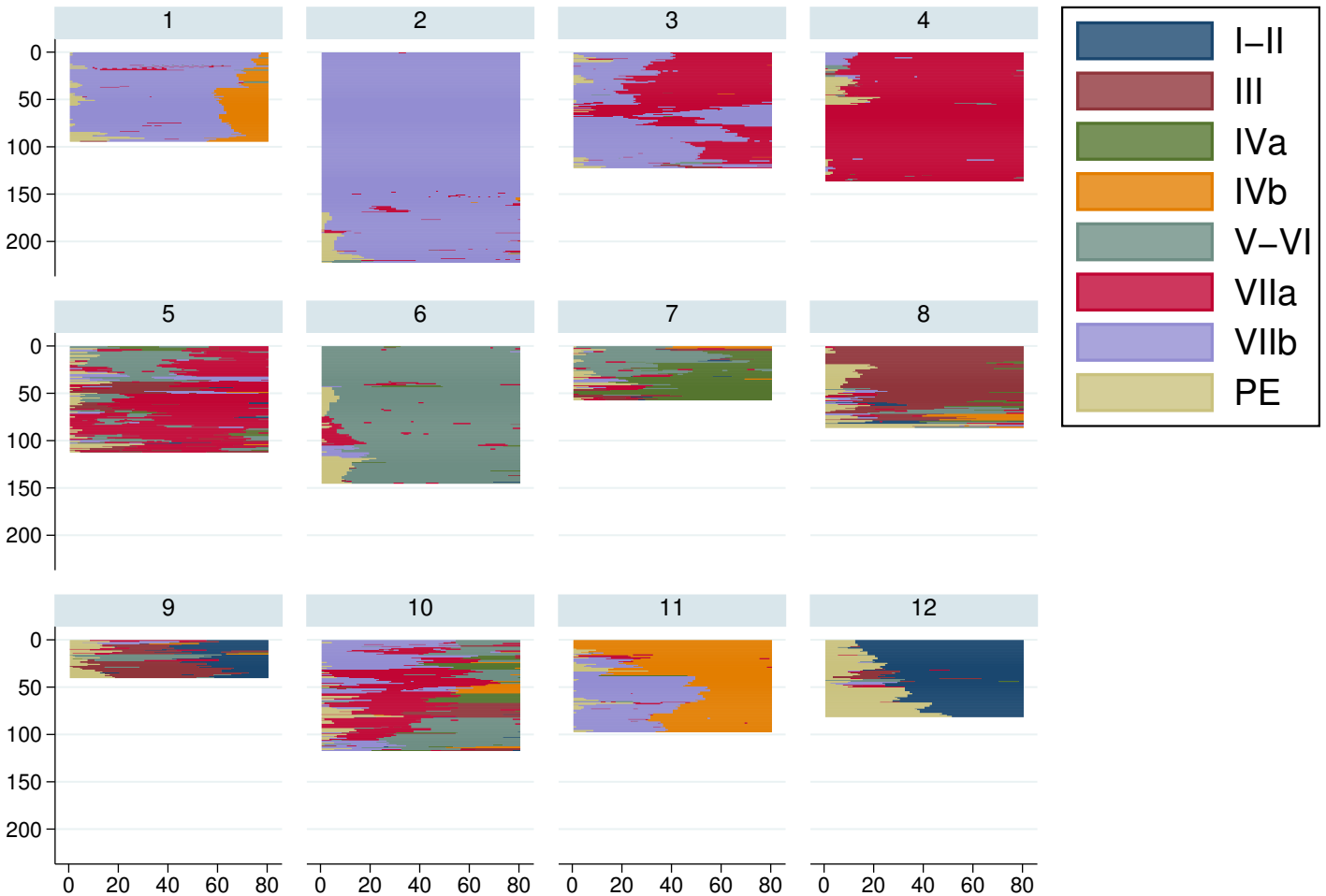
Code

```
matrix sub = (0.0, 2.0, 2.0, 2.0, 2.0, 3.0, 3.0, 1.5 \ ///
              2.0, 0.0, 1.0, 1.0, 1.0, 2.0, 2.0, 1.5 \ ///
              2.0, 1.0, 0.0, 1.0, 1.0, 2.0, 2.0, 1.5 \ ///
              2.0, 1.0, 1.0, 0.0, 1.0, 2.0, 2.0, 1.5 \ ///
              2.0, 1.0, 1.0, 1.0, 0.0, 2.0, 2.0, 1.5 \ ///
              3.0, 2.0, 2.0, 2.0, 2.0, 0.0, 1.0, 1.5 \ ///
              3.0, 2.0, 2.0, 2.0, 2.0, 1.0, 0.0, 1.5 \ ///
              1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 0.0)

oma s1-s80, subs(sub) indel(1.5) pwd(oma) len(80)

clustermat wards oma, add
cluster gen g = groups(4/12), ties(fewer)
cluster gen g999 = groups(1999), ties(fewer)
```

Chronogram by cluster



Indexplot by cluster



Graphs by g12

Criticisms

- Early: not sociological
- Intermediate: what do we need to do to make it sociological
- Late: Sequence analysis is just a step on the road to more holistic analyses of lifecourse data

Conclusion

- SA is good but not enough to displace GLR:
  - Abbott over-optimistic
- It has exploratory/descriptive strength but
  - a lot of "researcher degrees of freedom"
  - no stochastic framework
- It needs to be (and increasingly is being) related to other techniques
  - established techniques like hazard models
  - newer approaches like Hidden Markov Models

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

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

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

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 optimal matching methods in sociology. *Sociological Methods and Research*, 29(1):3–33.

Aisenbrey, S. and Fasang, A. E. (2010). New life for old ideas: The 'second wave' of sequence analysis – bringing the 'course' back into the life course. *Sociological Methods and Research*, 38(3):420–462.

Barban, N. and Billari, F. (2012). Classifying life course trajectories: A comparison of latent class and sequence analysis. *Journal of the Royal Statistical Society Series C*, 61(5):765–784.

Billari, F. C., Fürnkranz, J., and Prskawetz, A. (2006). Timing, sequencing and quantum of life course events: A machine learning approach. *European Journal of Population*, 22:37–65.

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

Blanchard, P., Bühlmann, F., and Gauthier, J.-A., editors (2014). *Advances in Sequence Analysis: Theory, Method, Applications*. Springer, Berlin.

Gabadinho, A. (2014). Package 'pst'. probabilistic suffix trees and variable length Markov chains. Technical report, CRAN.

Halpin, B. (2013). Sequence analysis. In Baxter, J., editor, *Oxford Bibliographies in Sociology*. Oxford University Press, New York.

Halpin, B. (2014). SADI: Sequence analysis tools for Stata. Working Paper WP2014-03, Dept of Sociology, University of Limerick, Ireland.

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.

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.

Lovaglio, P. G. and Mezzanzanica, M. (2013). Classification of longitudinal career paths. *Quality and Quantity*, 47(2):989–1008.

Stovel, K. (2001). Local sequential patterns: The structure of lynching in the Deep South, 1882–1930. *Social Forces*, 79(3):843–880.

Stovel, K., Savage, M., and Bearman, P. (1996). Ascription into achievement. *American Journal of Sociology*, 102(2):358–99.

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.

Wuerker, A. (1996). The changing careers of patients with chronic mental illness: A study of sequential patterns in mental health service utilization. *The Journal of Behavioral Health Services and Research*, 23(4):458–470.