"New Life for Old ideas: The Second Wave of Sequence Analysis Bringing the Course Back Into the Life Course" by Silke Aisenbrey and Anette E. Fasang

Presented by Daniela R. Urbina Julio

November 8, 2018

Sequence Analysis: An Overview

- Technique used in the field of life course sociology.
- Life course patterns rely on two core theoretical concepts: a discrete **"transition"** and a holistic **"trajectory."**
- To analyze transitions scholars usually rely on event history models. Over-emphasizing the study of isolated transitions while neglecting the study of trajectories.

Sequence Analysis: An Overview

What is a sequence?

- By sequence we mean an ordered list of elements. The elements of a sequence are events, drawn from a set of all possible events in a set of sequences, the universe of events (Abbott, 1995).
- Events in a sequence can be unique or they can repeat.
- Sequences can have dependence between their states. The most familiar examples of this are stochastic processes, in which the $n + 1th$ element of the sequence is some specified function of the nth or perhaps earlier elements.
- A sequence can be investigated either for itself or as an independent or dependent variable.

Sequence Analysis: An Overview

What do we gain with sequence analysis?

- Study of events in context and holistic life course patterns.
- Study of variation between life courses and differentiation as variation within one life course over time.
- Not based in any assumptions about the processes that generate the data. SA rooted in the tradition of algorithmic exploratory data analysis.
- Informative on subgroups that are not at risk of experiencing certain predefined transitions (critique to event history analysis).

Optimal Matching Analysis

- Statistical technique at the chore of SA.
- Optimal Matching is a **dissimilarity measure** between sequences originally proposed in the field of information theory and computer science.
- Later introduced in the social sciences by Andrew Abbott.
- In OM, the degree of dissimilarity between two sequences is determined by the least number of edit operations that are necessary to turn one sequence into the other (Lesnard 2006). Solves an optimization problem.
- Three kinds of edit operations are generally used: **insertion, deletion, and substitution.**

Optimal Matching Analysis

We can summarize the OM procedure in three steps:

1. Theoretical specification of state space and transformation costs of edit operations
2. OM algorithm to produce pairwise distances between subjects.
3. Multidimensional scaling or clustering. Usually Ward's (1963) agglomerative hierarchical cluster technique.

Transformation Costs

Costs attached to each **indel or substitution operation** to turn $S2$ into $S1$.

	Episodes			
	0	**1**	**2**	**3**
S1	A	B	C	D
S2	D	A	B	C

Transformation Costs

- Indel versus substitution operations emphasize two distinct dimensions of a sequence: **occurrence** of states versus their **timing**.
- **Substitutions**: Concerned with whether the same state occurs at the same time point in two sequences. Timing is preserved, while the event itself is approximated.
- **Indel:** Finding their longest common subsequence, whatever their location in the two sequences. Preserve the event, but distort time.

Transformation Costs

- Setting costs to these two types of operation determines the importance of timing over the preservation of the event.
- Generally indel operations are undesirable in the social sciences (Lesnard 2006, Aisenbrey and Fasang 2010). E.g. study if careers or school-to-work transitions.
- The output is a gigantic dissimilarity matrix between all sequences (individuals, and not variables).

Main Criticisms to the First Wave of SA (1980-1990

1. Basic opposition to the algorithmic modeling culture. A belief that a probabilistic base is essential to any approach to data.

2. Specific critiques to optimal matching procedures.

Criticism I: Link between Theory and Transformation Costs

Solutions

- **Destination states**. E.g. first faculty position.
- Caveat: How to quantify distances between events leading to the final state?
- **Data-based Transformation Costs**: Sample specific substitution costs based on the frequency of transitions from one state to another in a given data set (Rohwer 1997). Substitutions of states with high transition rates are less costly.

- Caveat: Useful if costs are independent of directionality—state A to state B or state B to state A.

Criticism I: Link between Theory and Transformation Costs

Solutions

- **Reference sequences**: Introducing references or "ideal" sequences and constructing distances against this specific baseline (Abbott and Hrycak 1990).
- Caveat: Results may be difficult to interpret when a similar distance to the ideal type covers very different substantive sequence patterns.

Criticism IV: Order and Timing of States Within Sequences

Problem

- Inability of OA to account for the direction of time and the order of states across sequences (Wu 2000).
- Timing: Transformation costs are the same at each time point of the sequence. E.g. Unemployment and age.
- Order: Transformation from one state to the next is assigned the same cost independent of the direction. E.g. Employment-Unemployment.

Criticism IV: Order and Timing of States Within Sequences

Solutions for Timing: Lesnard's dynamic Hamming measure

- Measure that allows time dependent substitution costs based on transition matrices. Time point sample-specific substitution costs.
- Can be applied only on sequences of equal length.

Criticism IV: Order and Timing of States Within Sequences

Lesnard's dynamic Hamming measure

$$s_t(a,b)=\begin{cases}4-\left[p\left(X_t=a\middle|X_{t-1}=b\right)+p\left(X_t=b\middle|X_{t-1}=a\right)+p\left(X_{t+1}=a\middle|X_t=b\right)+p\left(X_{t+1}=b\middle|X_t=a\right)\right] & \text{if } a\neq b\\ 0 & \text{otherwise}\end{cases}$$

Where *a* and *b* are events. Substitution costs oscillate between 4 and 0.

Criticism IV: Order and Timing of States Within Sequences

Non-alignment solutions: Dijkstra and Taris (DT) coefficients.

- The analysis of ordered pairs of sequence elements to determine sequence similarity.
- DT coefficients measure sequence similarity based on the number of moves of sequence elements necessary to turn one sequence into another in a three-step procedure.

Criticism IV: Order and Timing of States Within Sequences

Step 1: Remove elements that are not common.

$$(1) \quad H \quad UE \quad UE \quad E \quad E \quad E \quad H$$

$$(2) \quad I \quad I \quad H \quad H \quad H \quad E \quad E$$

Criticism IV: Order and Timing of States Within Sequences

Step 2: Reduce sequence to an equal number of common states.

$$(1) \quad H \quad E \quad E \quad E \quad H$$

$$(2) \quad H \quad H \quad H \quad E \quad E$$

Criticism IV: Order and Timing of States Within Sequences

Step 3: Minimal number of moves to generate same sequence/order.

$$(1) \quad H \quad E \quad E \quad H$$

$$(2) \quad H \quad H \quad E \quad E$$

The total number of moves made in all three steps indicates agreement between both reduced sequences: The more moves, the higher the distance between them.

Criticism IV: Order and Timing of States Within Sequences

Non-alignment solutions: Dijkstra and Taris (DT) coefficients

- Caveats: DT coefficients have been criticized primarily for the discarding of states in the first and second step of reduction.

- Sequences without repetition of events.
- Disconnected from social theory.

Criticism IV: Order and Timing of States Within Sequences

Non-alignment solutions: Elzinga's Sequence Complexity

- Measure of variability within individual sequences over time.
- Number of distinct states, the order of states, and the variance of durations spent in different states.

Criticism IV: Order and Timing of States Within Sequences

Elzinga's Sequence Complexity Without Duration Variance

$$0 \leq C(x^n) = \log_2 \phi(x^n) \leq n$$

- Where complexity $C(x)$ of an n-long sequence x increases with the number of distinct subsequences.

Criticism IV: Order and Timing of States Within Sequences

Elzinga's Sequence Complexity With Duration Variance

$$C(x, t_x) = \log_2(\phi(x).T(x, t_x)) \text{ with}$$

$$1 \leq T(x,t_x) = \frac{V_{\max}(t_x) - V_{\min}(t_x) + 1}{V(t_x) - V_{\min}(t_x) + 1}$$

- Where T is the "relative variance inverted," and $V_T x$ denotes the variance of all subsequence durations.
- V_{min} and V_{max} denote the upper and the lower bound of duration variance of all subsequences, respectively.

Empirical Application Take-Away

- Empirical application—school-work transitions—OM and Hamming measure do a better job than non-alignment techniques.
- In the latter, 50% of cases were classified in the discontinuity or residual cluster.
- Sequence Complexity varies according to the SA technique.

Newer Developments

Multichannel Sequence Analysis: Gauthier et al.

- Extends OM to describe individual trajectories on several dimensions simultaneously. Example: Life-Work trajectories.
- This approach considers two or more channels per individual and uses one cost matrix for each channel.
- Channels associated with a given individual are synchronized so that, for example, the *xth* character of channel A and the *yth* character of channel B correspond to the same year for a given individual.
- Eventually, the contribution of channels A and B is averaged to yield the final cost associated with the matching of positions x and j for the two individuals.

Newer Developments

Sequence Analysis Multistate Model Procedure: Studer, Struffolino and Fasang.

- Combines long-run processes (sequence analysis) with time varying covariates that explain the entrance to specified states or transitions.
- Marriage between SA and event history models.
- Example: How (1) time-varying statuses in the family domain are associated with women's employment trajectories in East and West Germany.

Newer Developments

Sequence Analysis Multistate Model Procedure: Studer, Struffolino and Fasang.

- **First step:** Use SA to extract subsequences that start with a transition and lasts for t time units. Subsequences describe the transitions between two states as well as what follows over a period of t time units.

Newer Developments

Sequence Analysis Multistate Model Procedure: Studer, Struffolino and Fasang.

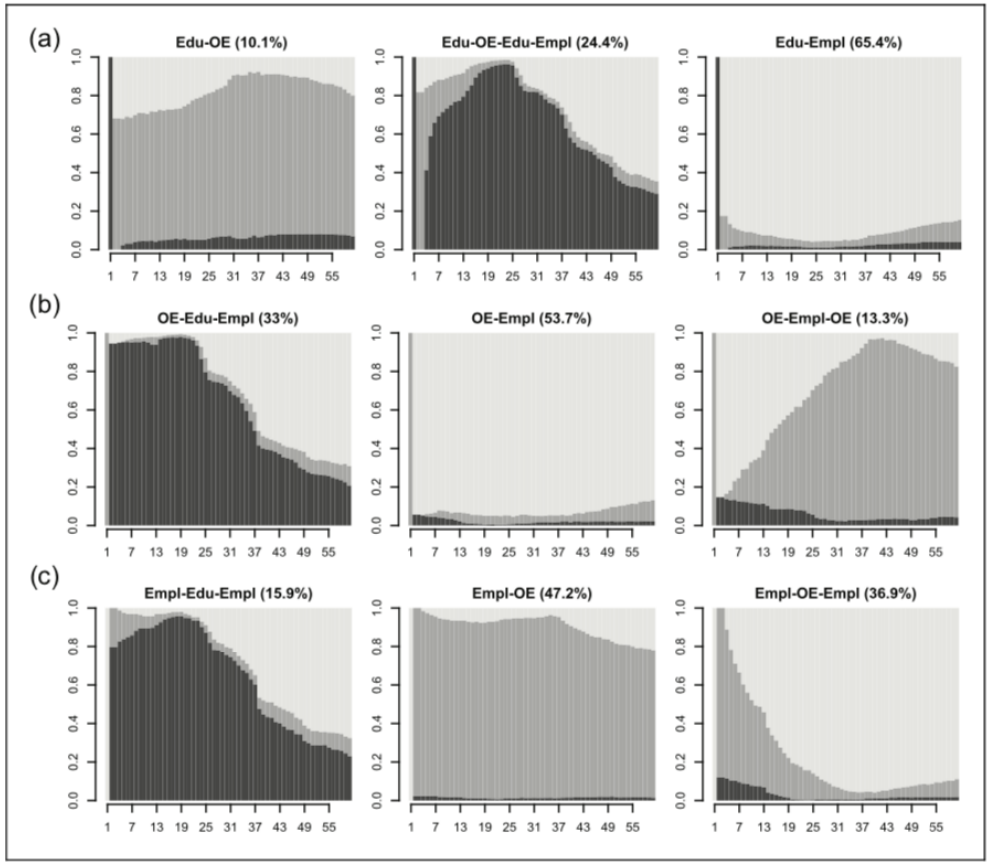


Figure 3. Clusters of subsequences: distribution of the states (y-axis) at each month following a transition (x-axis). Percentages for each plot represent the proportion of subsequences in each cluster departing from a specific state.
Source: NEPS data.

Newer Developments

Sequence Analysis Multistate Model Procedure: Studer, Struffolino and Fasang.

Second step: Multistate event history models to assess how covariates are associated with the hazard rate of following each type of subsequence cluster while departing from a given state.Example: Covariates such as being in an union and having at least one child.

Thank you!

Criticism 2: Validation of Results

SA criticized for "fishing for patterns" and lacking standards for validation (Wu 2000).

Problems and Solutions

- **Choose cluster algorithm technique**.
- Solution: Check robustness of findings under different cluster algorithms (Martin et al 2008).

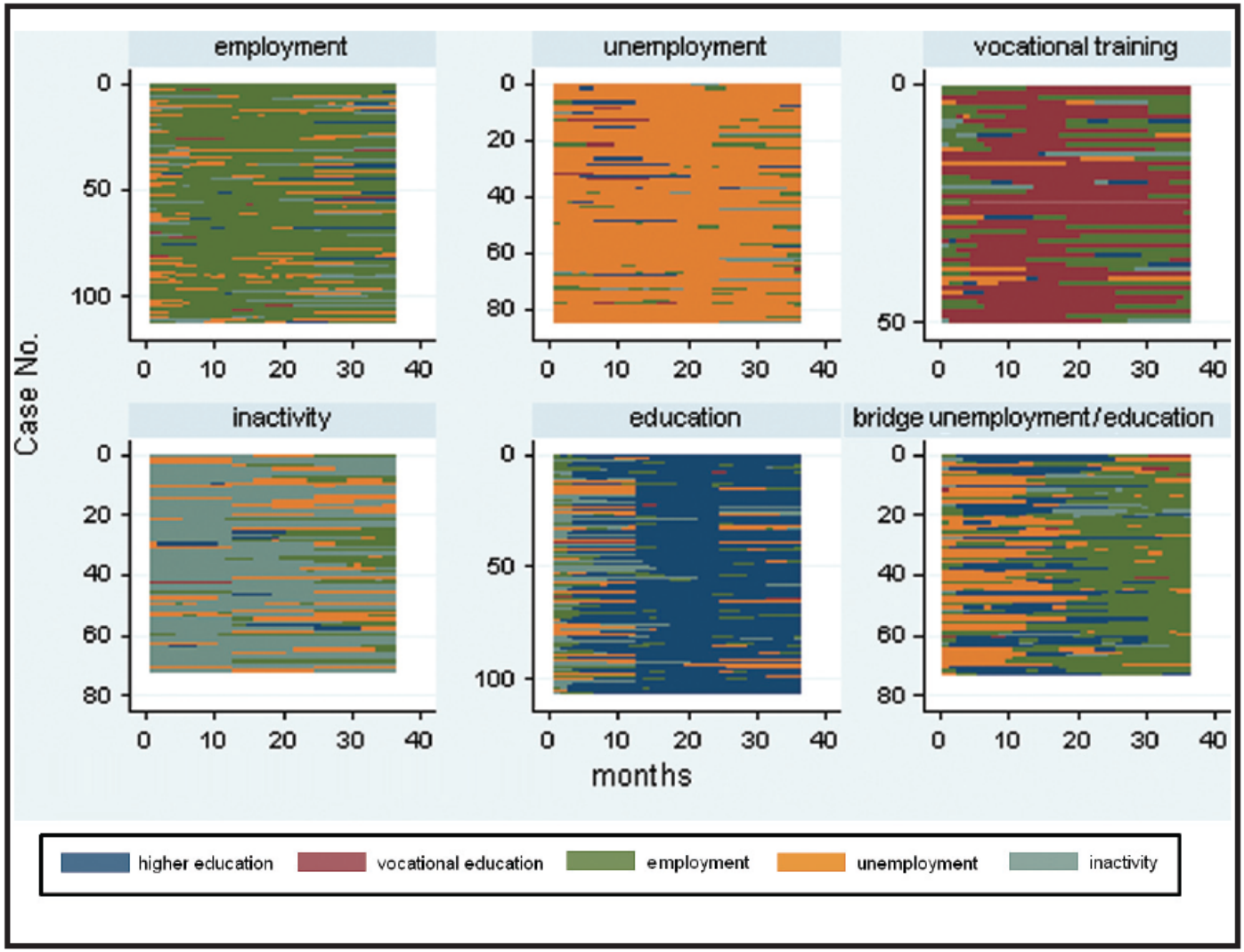
Criticism 2: Validation of Results

Problems and Solutions

- **Determine number of clusters.** Traditional cluster cutoff criteria are not transferable for the SA case, because distance matrix is based on sequence comparison and not on some distance measure (e.g. gap statistic).
- Solution: Comparing mean within-cluster distances to mean between-cluster distances for a range of cluster solutions (Calinski and Harabasz 1974; Milligan and Cooper 1985).

Criticism 2: Validation of Results

Problems and Solutions



Criticism 3: Missing and Incomplete Data

Problems and Solutions

- **Unequal sequence length due to censoring (Wu 2000).**.How to minimize sequence dissimilarity due to unequal sequences caused by restrictions of the observation window? (Not related to the process of interest)
- Variable indel costs (Stoven and Bolan 2004). Reduce indel costs in the case of sequences of unequal length.