COMPLEX TRAJECTORIES

Sequence Analysis

May 2023

Co-funded by the Erasmus+ Programme of the European Union

Author:

Janine Jongbloed

The following documents are embedded in a whole MOOC course on Longitudinal analysis techniques developed in the framework of the same Complex Trajectories project. Those interested in doing the MOOC should consider the following procedure:

1. Access the AULAbERTA space through the link: https://aulaberta.uab.pt/

2. Register in the platform a. Select one of the available languages (Portuguese or English) b. Follow the instructions given in the platform

3. After creating the account, they should access the MOOC Longitudinal Analysis through the link: https://aulaberta.uab.pt/course/view.php?id=94

Co-funded by the Erasmus+ Programme of the European Union

Sequence analysis of complex trajectories

Janine Jongbloed, IREDU, Université de Bourgogne

Co-funded by the Erasmus+ Programme of the European Union

Hello and welcome to unit 3, focused on sequence analysis.

Outline

Brief history of sequence analysis

How it works

Typical methodological steps

Challenges

Co-funded by the Erasmus+ Programme of the European Union

In this presentation, I'll begin with a brief history of sequence analysis in the social sciences, and then describe how it works. In a second presentation, I'll outline the typical methodological steps taken when we use sequence analysis and some of the challenges that we might face.

Brief history

How sequence analysis became a sociological tool

Co-funded by the
Erasmus+ Programme
of the European Union

First, we begin with some historical context.

Roots in biology

- Sequence analysis (SA) with optimal matching (OM) is a set of methods borrowed from the field of molecular biology where they're used to analyze proteins and DNA
- Adoption in sociology driven mainly by American sociologist Andrew Abbott
  - Abbott and Forrest (1986): patterns in order of steps in traditional dances in rural England in the nineteenth century
- Collection of longitudinal analysis techniques used to address a variety of research topics:
  - careers (employment biographies)
  - life course (education, work, family formation)
  - cultural symbols (dances, folktales)
  - conversations, articles…

Co-funded by the Erasmus+ Programme of the European Union

Sequence analysis is a type of descriptive quantitative data analysis technique taken from biology, where it's used to study proteins and DNA. This approach was adopted in the social sciences for the first time by the sociologist Andrew Abbott, who used a novel application of this type of analysis to study patterns in the order of dance steps in historical traditional dances in England.
Since these initial applications, sequence analysis has been applied to many different topics, including career and retirement patterns, patterns in work and family formation, school-to-works transitions of youth, and many more.

Sequence questions

- Sequence methods allow us to answer questions about "whether some process or series of events typically happens in a particular order" (Abbott & Tsay, 2000)
- Optimal matching (OM) and classification techniques allow us to further consider:
  1. "Are there patterns among a set of sequences?"
  2. "If such patterns exist, how are they produced?"
  3. "What are the consequences of such patterns?" (Abbott, 1990)

Co-funded by the Erasmus+ Programme of the European Union

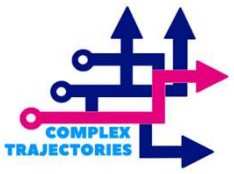
Sequence analysis in the social sciences obviously answers different questions than it does in the biological sciences. In our application of sequence analysis, we use these methods to answer questions about "whether some process or series of events typically happens in a particular order". This allows us to consider if there are patterns in a set of sequences, what variables are linked to producing these types of patterns, and what the consequences of these patterns are at the end of the sequence.

Research aims

- Most often exploratory and descriptive
  - allows use to identify "patterns of social processes over time" (Aisenbrey & Fasang, 2010, p.432)
- Takes an event-focused, "narrative" approach to sociology (Halpin, 2013)
- Typically, NOT concerned with hypothesis testing
  - but may be "hypothesis-generating"
  - need further analytical steps to answer questions about what explains the sequences and what they in turn explain

Co-funded by the Erasmus+ Programme of the European Union

Thus, our research aims when using sequence analysis are typically exploratory and descriptive and focused on finding patterns in longitudinal data. In doing so, we focus on a storyline of events in a « narrative » pattern-finding approach, that can then be integrated into further types of analysis that test different variables linked to these patterns.



How it works

Sequence analysis with optimal matching & clustering – explained in simple terms!

Co-funded by the
Erasmus+ Programme
of the European Union

Next, we will explore the steps involved in a typical methodological approach using sequence analysis in the social sciences.

Sequence analysis

- Involves ordering a list of elements across time into sequences (ordered arrays)
  - we meaningful code the (categorical) data over some period measured at regular intervals
  - positions of the elements are fixed and ordered by elapsed time
- Focuses on sequences as wholes rather than discrete "transitions"
- In sociology these elements are coded as states in successive time-periods
  - for example, life-histories or employment biographies
  - states may include education, employment statuses, marital statuses, residences, family formation…

Co-funded by the Erasmus+ Programme of the European Union

Our first step is to code our data into an ordered array. A typical example is to code an activity status or state each month or year over a period of time. The activity is a categorical variable that can take on a limited number of values. For example, we might look at whether someone is employed full-time or part-time, in education, or unemployed each month. We would then have a series of states for each individual, forming one whole sequence.

Complex Trajectories

"Mosaic of Post-Secondary Attendance"
(Andres & Offerhaus, 2012)

FIGURE 2: Institutional States

- Not attended
- Community college
- University college
- Technical or vocational training institute
- University
- Private institute
- Combination of several institutes
- Other

FIGURE 4: Post-Secondary Institutional Attendance, 1988-2010

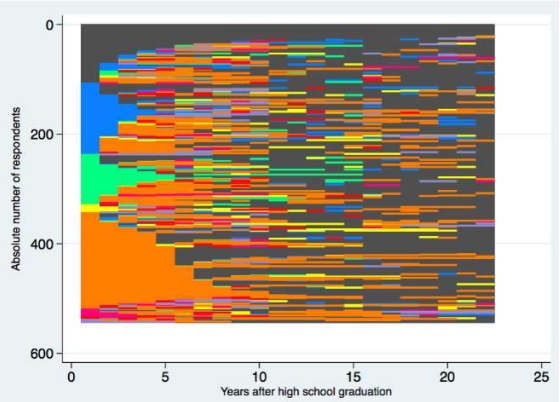


FIGURE 3: Three Examples of Trajectories Over 22 Years

Co-funded by the Erasmus+ Programme of the European Union

9

For example, here we see sequences of post-secondary education attendance in British Columbia, Canada. Each line on the two figures shows one individual's whole sequence of yearly participation across 22 years. We see a great variety of sequences, a mosaic as the authors describe it, that are made up of individual categorical data points organized in an ordered array.

Optimal matching (OM)

- OM is an algorithm that creates a metric for pairwise distances between whole sequences
- We measure distances using elementary operations that turn one sequence into another
- We must assign costs to the elementary operations
- Sequences can be changed in two basic ways:
  1. we can replace (or substitute) an element with a different one
  2. we can insert or delete an element from the sequence ("*indel*")

For more detail, refer to: Abbott, Andrew, and Angela Tsay. 2000. "Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect." *Sociological Methods & Research* 29 (1): 3–33. doi:10.1177/0049124100029001001.

Co-funded by the Erasmus+ Programme of the European Union

To find meaning in the variety of sequences that we find, we can use a second group of statistical techniques to compare all possible pairs of individual sequences. The most often used technique in the social sciences is optimal matching. This is a way of deciding how similar or dissimilar sequences are from other sequences and computing these scores across all pairs of individual sequences. We can do this by either substituting one state with another state, or by inserting or deleting an extra element in the sequence.

Costs

- Need to determine costs:
  - we specify replacement or substitution costs with a simple matrix
    - insertion and deletion costs are set to a given value
    - we can make some substitutions less costly than the combination of one deletion followed by one insertion
    - this is not sensitive to the direction of a transition!
  - we can also use observed transition rates, where we assume that the frequency of transitions between states gives information about the extent of similarity between these states (Dlouhy & Biemann, 2015)
    - costs are inversely proportional to transition frequencies
    - an "objective" approach

Co-funded by the Erasmus+ Programme of the European Union

When we do this, we aim to find the least costly way possible to change one of the individual sequences into the other. We can do so by setting a theoretically-driven substitution matrix that weights some changes more strongly than others by giving them a higher cost. We can also use a data-driven approach, where we use the observed transition rates to inversely weight changes to the transition frequencies that we see in the data. This makes less common changes or transitions more costly.

Distances between sequences

- All sequence pairs are "matched" with one another to create a distance matrix
  - pairwise distance or dissimilarity scores are determined through the creation of substitution, insertion and deletion costs
  - used to compute the "cost" of transforming one sequence into another, which is done for all sequence pairs
  - difference between any two sequences is the "cost" of the "cheapest" combination of substitutions and insertions or deletions that change one sequence into the other (Halpin, 2013)
  - proximity = belong to similar social rhythms (Lesnard, 2014)

For more detail, refer to: MacIndoe, Heather, and Andrew Abbott. 2012. "Sequence Analysis and Optimal Matching Techniques for Social Science Data." *Handbook of Data Analysis*, 386–406. doi:10.4135/9781848608184.n17

Co-funded by the Erasmus+ Programme of the European Union

After using one of these techniques, we end up with a matrix of measures of the distances between each pair of individual sequences in the dataset. Greater distances show that two sequences are more dissimilar from one another. These distances are always the smallest costs necessary to change one sequence into the other. Those who have small distances from one another can be seen as belonging to similar social rhythms or patterns.

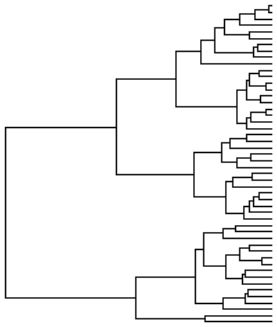
Cluster analytic techniques

- Synthesizes complex information, data simplification
- Takes the large, heterogeneous group of sequences and divides it into smaller, homogeneous groups
  - these smaller groups are similar to within-group members and different from other groups
- Many different clustering algorithms
  - depends on one's purpose in clustering, the type of input variables
  - different methods may produce different results
- Hierarchical agglomerative methods
  - use a proximity matrix to join the two most similar cases, then the next two cases with the smallest distance and so on for N − 1 iterations
  - most popular approach is Ward's method, which reaches the minimum variance, or error sum of squares (ESS), within clusters

Co-funded by the Erasmus+ Programme of the European Union

Finally, we can use this matrix of distance measures and find patterns within the sequences. To do this, we use some kind of grouping technique, and most often cluster analysis techniques. There are a number of different clustering techniques depending on our goals and data, but all group similar sequences together into clusters. The most popular is Ward's method, which minimizes increases in the total within-cluster variance after merging each time that cases are added. So sequences that are more similar to each other will end up in the same group.

Clustering



- Examining the dendrogram (tree diagram) of possible number of groupings in the data
  - find best fit in terms of statistical differentiation and theoretical interpretability of the group characteristics (e.g., Cramer Ridder tests)
- We use pairwise distances output from the OM algorithm as data input
- We typically generate data-driven typologies (Halpin, 2013), showing patterns or ‘‘sequential equivalence” in trajectories (Han & Moen,1999)
- Resulting typology can be used as a dependent or independent variable in further analyses

Co-funded by the Erasmus+ Programme of the European Union

We will end up with a number of different options for the number of cluster groups to choose. Here, we must decide using both theoretical and statistical evidence what number of groups best illustrates the existing patterns in the sequence data. We can use both descriptive statistics and visualization techniques to examine different numbers of clusters, as well as different objective measures.

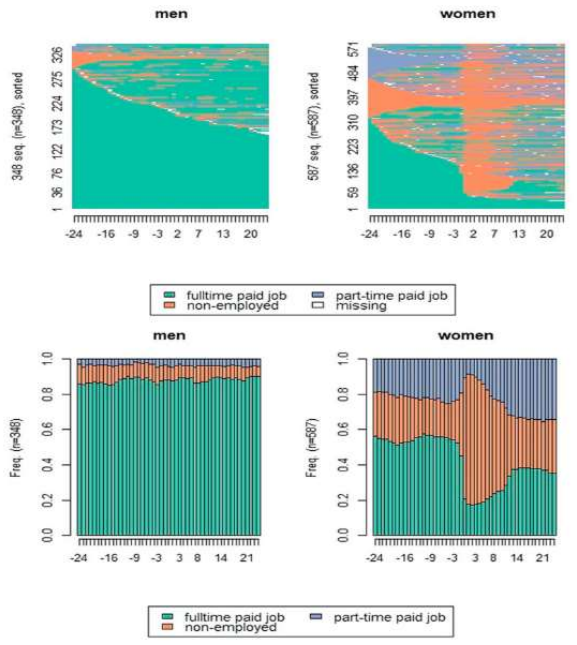


Figure 5. Employment sequences of parents in BC surrounding childbirth (first row) and monthly frequencies (second row).

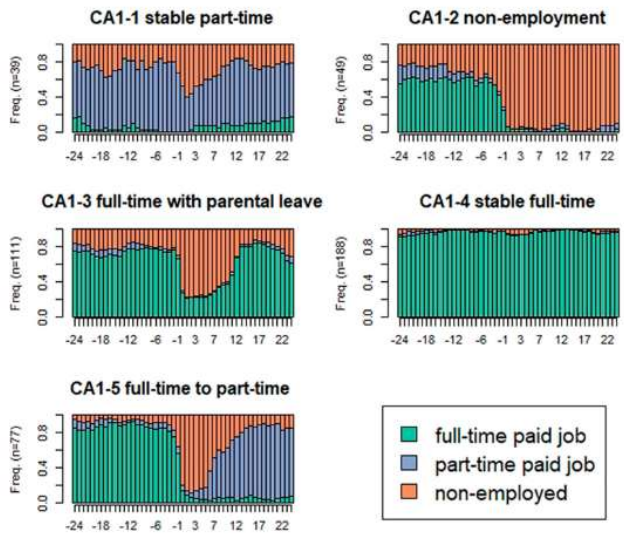


Figure 6. The employment sequences of parents in BC surrounding the birth of a first child.

(Antonini, Pullman, Fuller & Andres, 2020)

Co-funded by the Erasmus+ Programme of the European Union

15

For example, here on the left we see the state distribution plots of a sample of adults showing their employment histories. Each line no longer corresponds to one individual sequence but rather, for the whole sample, the proportion of individuals at each different state for a different period in time.
On the right, we see five different cluster groupings that emerged in the data. In our next video, we'll summarize what we've covered so far in five simple steps, then talk in more detail about each of these types of visualization techniques, and then the analyses that we can use to predict cluster membership and see the consequences of different sequence patterns, as well.

Thank you!

Please refer to the supporting materials on the MOOC website.

16

Thank you, and see you in the next video!

Sequence analysis of complex trajectories

Janine Jongbloed, IREDU, Université de Bourgogne

Co-funded by the Erasmus+ Programme of the European Union

Hello and welcome back to unit 3, focused on sequence analysis.

Outline

Brief history of sequence analysis

How it works

Typical methodological steps

Challenges

Co-funded by the Erasmus+ Programme of the European Union

We've already talked about the history of sequence analysis in the social sciences, and described how it works. Now, I'll outline the typical methodological steps taken when we use sequence analysis to answer a particular research question and some of the challenges that we might face when doing so.

5 simple steps

Sequence analysis as a research methodology…

Co-funded by the
Erasmus+ Programme
of the European Union

The sequence analysis methodology we've already discussed can be summarized in five simple steps.

Typical 5-step approach

(1) Describe the sequences

(2) Visualize the sequences
- e.g., sequence index plots

(3) Compare sequences
- e.g., optimal matching

(4) Group sequences
- e.g., clustering techniques

(5) Associate groups with other variables
- e.g., regression analyses

(Mills, 2014)

Co-funded by the Erasmus+ Programme of the European Union

These steps are taken from a book on the subject, written by Melinda Mills, but have also been employed by many researchers in the field. I list the steps here, and then we'll describe each one in more detail. The « typical pattern » in sequence analysis methodologies is to describe the sequences using descriptive statistics, visualize the sequences, compare them with one another, group them, and then look at how they relate to other variables that are measured before or after the sequence under question.

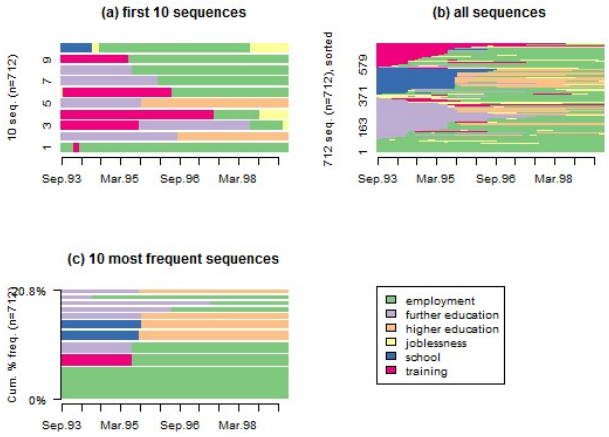
(1) Describe the sequences

- We often use aggregated indicators:
  - average duration of states
    - sum of the total number of months spent in one state (consecutive or not), frequency of states (Brzinsky-Fay, 2007)
  - average number of episodes in specific states
  - average number of episodes overall
    - might be seen as showing flexibility, complexity, disorder
    - can also assess the level of volatility (Brzinsky-Fay, 2007)
- Can compare these across pre-existing groupings in the data, such as gender, country, social origin…

Co-funded by the Erasmus+ Programme of the European Union

So our first step is to describe the sequences. We can use a number of different descriptive statistical measures, such as the average duration in different states, either the total time spend on each state or the average number of episodes (which are consecutive) in different states. We can also look at the number of different episodes and their durations as a measure of complexity. One option at this stage is to compare sequences across different pre-existing groups in our data, such as gender or socio-economic status.

(2) Visualize the sequences



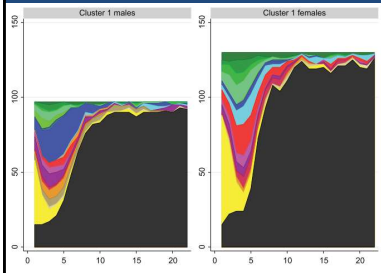
- index plots of individual or all sequences
- most frequent sequences
- state distribution plots showing the proportion of individuals in each state at each point in time…

Gabadinho, A., Ritschard, G., Müller, N.S. & Studer, M. (2011), Analyzing and visualizing state sequences in R with TraMineR, *Journal of Statistical Software*. Vol. 40(4), pp. 1-37.

Co-funded by the Erasmus+ Programme of the European Union

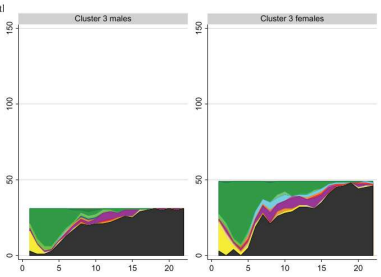
Next, we typically visualize the data using a number of existing tools. In R, we can use the TraMineR package to visualize our data. We show some or all individual sequences in an index plot, which has time on the horizontal axis and shows a line for each individual's sequence. We can also show the most frequently occurring sequences by their frequency in the dataset.

(2) Visualize the sequences



Cluster 1. The High Engagement Stratified Patl

"Applied and general higher education gender stratification in Canada" (Pullman & Andres, 2015)



Cluster 3. The Social Science to Applied Pathway.

Table 1. Distribution of Sample by Type of Programme, Total Years in Programme and Gender.

Type	Programme	Colour	Total Year	Total %	Male Year	Male %	Female Year	Female %
General	Fine arts		133	3.34	34	2.24	99	4.01
	Social sciences		509	12.79	184	12.12	325	13.18
	Humanities		116	2.91	30	1.98	86	3.49
	Life sciences		272	6.83	107	7.05	165	6.69
	Physical sciences		72	1.81	54	3.56	18	0.73
	Undeclared		667	16.68	218	14.36	449	18.21
Applied	Computer science		63	1.58	54	3.56	9	0.36
	Engineering		252	6.33	207	13.64	45	1.82
	Education		409	10.27	86	5.76	323	13.10
	Law		28	0.70	15	0.99	13	0.53
	Health		459	11.53	128	8.43	331	13.42
	Fitness and kinesiology		88	2.21	26	1.71	62	2.51
	Business		494	12.41	191	12.58	303	12.29
	Applied social science[a]		161	4.04	24	1.58	137	5.56
	Natural Resources		87	2.19	53	3.49	34	1.38
	Trades		81	2.03	78	5.14	3	0.12
	Administration		45	1.13	12	0.79	33	1.34
	Service		48	1.21	17	1.12	31	1.26
	Total		**3,984**		**1,518**		**2,466**	

[a]Applied social science includes journalism, social work, architecture and library sciences.

Co-funded by the Erasmus+ Programme of the European Union

We can also show our data using a state distribution plot, where the proportion of individuals in each state at each point in time are shown. Using this approach, we no longer see individual sequences, but rather trends across individuals in the whole sample or group. Here, we see different fields of education and the proportion of individuals enrolled in each field of study at each point in time.

(3) Compare sequences

- OM is an algorithm that creates a metric for pairwise distances between whole sequences
- We measure distances using elementary operations that turn one sequence into another
- We must assign costs to the elementary operations
- Sequences can be changed in two basic ways:
  1. we can replace (or substitute) an element with a different one
  2. we can insert or delete an element from the sequence ("*indel*")

For more detail, refer to: Abbott, Andrew, and Angela Tsay. 2000. "Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect." *Sociological Methods & Research* 29 (1): 3–33. doi:10.1177/0049124100029001001.

Co-funded by the Erasmus+ Programme of the European Union

In step 3, we compare sequences using a chosen statistical approach, such as optimal matching, which we discussed in the last video. This is a way of deciding how similar or dissimilar sequences are from other sequences and computing these scores across all pairs of individual sequences.

(3) Compare sequences

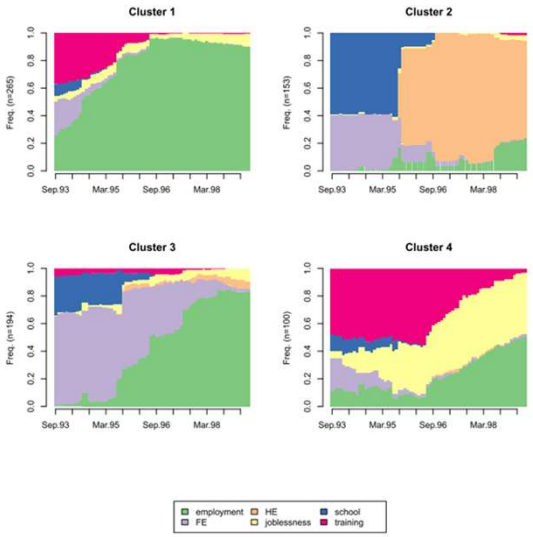
- Typical optimal matching procedure:
  1. theoretical specification of state space and transformation costs,
  2. optimal matching algorithm to produce pairwise distances between subjects (Aisenbrey & Fasang, 2010)
- Importance of cost setting:
  - substitutions emphasize the timing of states
  - *indel* operations emphasize the occurrence of states
  - if focus is on timing and order of events, make *indel* costs high (MacIndoe & Abbott 2004)
  - can use more complex methods, e.g., 'dynamic Hamming distance' is based on time varying substitution costs (Barban & Billari, 2012)

Co-funded by the Erasmus+ Programme of the European Union

As described previously, we need to make a number of decisions that will impact the resulting scores, based on our research questions and theoretical approaches. Notably, we'll choose between a self-constructed substitution matrix based on theory and a data-driven matrix based on transition probabilities.

(4) Group sequences

- multidimensional scaling or clustering procedures
- expected outcome is a typology of cluster groupings of similar trajectories
- may compare to theoretical sequence types



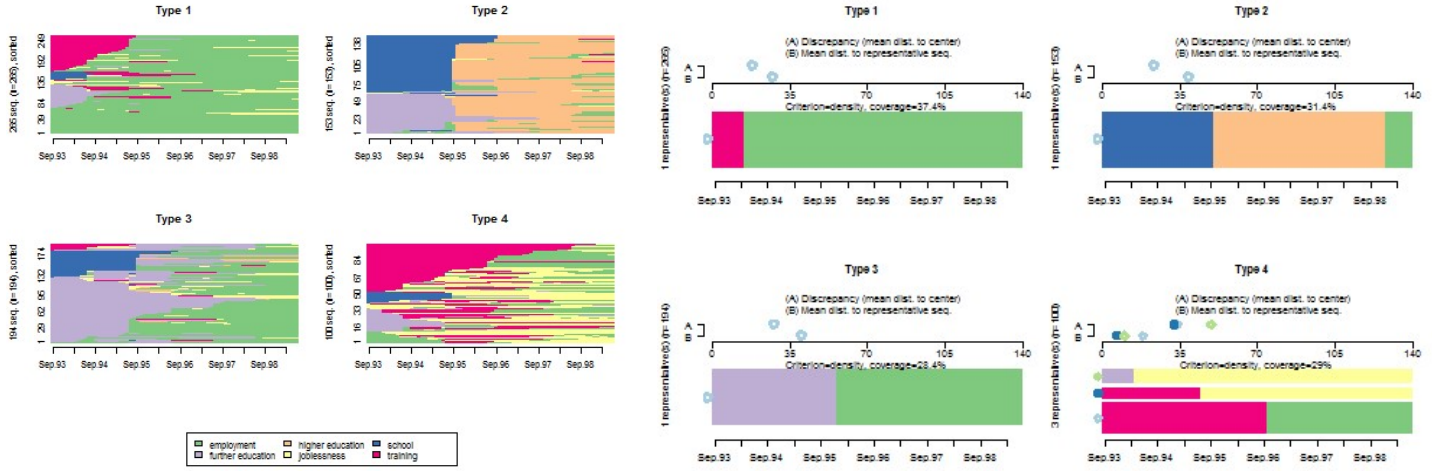
Gabadinho, A., Ritschard, G., Müller, N.S. & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR, *Journal of Statistical Software*. Vol. 40(4), pp. 1-37

http://traminer.unige.ch/

Co-funded by the Erasmus+ Programme of the European Union

Next, based on our work in step 3, we group sequences that are similar to one another. Typically, we use cluster analysis procedures to do this, and create a typology of clusters. There are many different clustering techniques, and your choice will depend on your own data and research questions. Useful information can be found on the TraMineR website shown here.

(4) Group sequences



Co-funded by the Erasmus+ Programme of the European Union

Gabadinho, A. & Ritschard, G. (2013), "Searching for typical life trajectories applied to childbirth histories", In Levy, R. & Widmer, E. (eds) Gendered life courses - Between individualization and standardization. A European approach applied to Switzerland, pp. 287-312. Vienna: LIT.

Once we've grouped sequences, we can then visualize the sequences within each group to illustrate patterns. Here we see different clusters that show different patterns in education, employment and joblessness. We see clearly that the dominant states at different points in time differ between groups, with a large proportion of individuals in each cluster showing the same state at similar points in time. This can be illustrated both by state distribution plots and by representative sequences for each cluster.

(5) Associate groups with other variables

- What determines trajectory types?
  - Groups may then serve as dependent variables in analyses predicting the impact of various (pre-existing) characteristics on membership
  - e.g., multinomial regression analyses
- What outcomes are these trajectory types associated with?
  - Can use various regression techniques to predict end-of-trajectory (or later) variables
  - e.g., educational attainment, income, well-being…

Co-funded by the Erasmus+ Programme of the European Union

Finally, we can associate our groups of sequences with other variables, to find what explains group membership and what the consequences are for those who follow a particular pattern. To do so, we often use different regression techniques. Since we have variables that represent our clusters, and for each individual whether their sequence falls into a particular cluster, we can then use multinomial regression to predict group membership in these clusters. And we can use ordinary least squares or other types of regression to predict such outcomes as being in employment, income, or job satisfaction. We can include our cluster variables as predictor variables for these different outcomes.

Challenges

Potential difficulties and drawbacks of a sequence analytical approach

Co-funded by the
Erasmus+ Programme
of the European Union

While simple in theory, many challenges might arise while conducting these analyses.

Difficulty 1

- We cannot model the social processes that generate the sequences—we need other methods for that (Halpin, 2013)
  - Rather we can describe and explore the overall structure of complicated longitudinal data, provide *context* for other models (Halpin, 2009)

Co-funded by the Erasmus+ Programme of the European Union

One challenge is mainly theoretical: Since we cannot model social processes per se using this approach, we are limited to what is in essence more of a descriptive analytical approach, as I emphasized in the first video. So it is important to take this into consideration as we interpret our results.

Difficulty 2

- A potential drawback is that we typically require long observation periods in order to gain meaningful insights from the OM and clustering stages
- Monte Carlo simulations of published OM studies have led to the recommendation of a minimum sequence length of 25 elements, assuming that there is variability in the data (Dlouhy & Biemann, 2015)
  - One problem that may result from small state space is tied distances (Martin et al. 2008)

Co-funded by the Erasmus+ Programme of the European Union

Another challenge is that we typical require data over relatively long (or frequently measured) periods of time in the sequence data to uncover meaningful patterns. Some simulation studies have suggested that we need around 20 or 25 points in time. Depending on our research question, this might mean that we need to devote a long time to data collection over several or many years. This approach also works best with data that have a large amount of variability, or a large number of different unique sequences.

Difficulty 3

- We need sequences of more or less the same length
- May need to standardize the pairwise distances if sequence lengths differ too much amongst sequences, as this can bias results (MacIndoe & Abbott, 2012)
- Recommended to use only sequences that have less than 30% elements missing & where the length of the shortest sequences is at least 70% of the longest sequences (Dlouhy & Biemann, 2015)

✓ *However, many creative technical solutions exist for various data challenges… e.g., see Aisenbrey & Fasang, 2010, p. 427!*

Co-funded by the Erasmus+ Programme of the European Union

Finally, a further challenge is that we need sequences of more or less the same length to avoid biasing our results. However, these and other data challenges can be addressed with some creative statistical and analytical approaches described in the literature (you can see one such reference here).

Thank you!

Please refer to the supporting materials on the MOOC website.

17

Thank you for your attention, and please refer to the supporting materials on the MOOC website! See you in the next unit.