INTRODUCTION TO SEQUENCE ANALYSIS

Luca Badolato*

Summer Incubator Workshops Series 2022

Max Planck Institute for Demographic Research

July 26, Rostock

*Department of Sociology and
Institute for Population Research,
The Ohio State University

Intro to Sequence Analysis. Goals of the workshop:

- Provide an introduction to Sequence Analysis:
  - Purposes of Sequence Analysis, object of analysis, methodological and historical developments across the years
  - Getting familiar with the basic measures and methods
  - Visualizing sequences and computing sequence indicators
- Show an application of Sequence Analysis in R using the package TraMineR

Sequence Analysis, where to find information

Sequence Analysis Association

HOME RESOURCES EVENTS COMMITTEE CONTACT LOGIN

Welcome to SAA

The Sequence Analysis Association (SAA) aims to promote research, teaching and diffusion of sequence analysis (SA) and its relationships with related methods.

To this end, the SAA will among others organize events such as symposiums, webinars, and training courses, collect and share information on SA related events, provide links to SA resources.

The SAA was created during the International Symposium on Sequence Analysis and Related Methods held on the 10-12 October 2018 at Monte Verità, TI, Switzerland.

Don't know what sequence analysis is about? See Wikipedia.

News

Webinar Recording: Studying Migration Using Sequence Analysis (28th April 2022) [...]

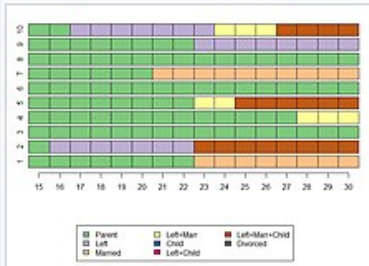
New Wikipedia page on SA in social sciences [...]

Sequence Analysis, where to find information

Sequence analysis in social sciences

From Wikipedia, the free encyclopedia

In social sciences, **sequence analysis (SA)** is concerned with the analysis of sets of categorical sequences that typically describe longitudinal data. Analyzed sequences are encoded representations of, for example, individual life trajectories such as family formation, school to work transitions, working careers, but they may also describe daily or weekly time use or represent the evolution of observed or self-reported health, of political behaviors, or the development stages of organizations. Such sequences are chronologically ordered unlike words or DNA sequences for example.



Index plot of 10 family life sequences

SA is a longitudinal analysis approach that is holistic in the sense that it considers each sequence as a whole. SA is essentially exploratory. Broadly, SA provides a comprehensible overall picture of sets of sequences with the objective of characterizing the structure of the set of sequences, finding the salient characteristics of groups, identifying typical paths, comparing groups, and more generally studying how the sequences are related to covariates such as sex, birth cohort, or social origin.

Introduced in the social sciences in the 80s by Andrew Abbott,[1][2] SA has gained much popularity after the release of dedicated software such as the SQ[3] and SADI[4] addons for Stata and the TraMineR R package[5] with its companions TraMineRextras[6] and WeightedCluster.[7]

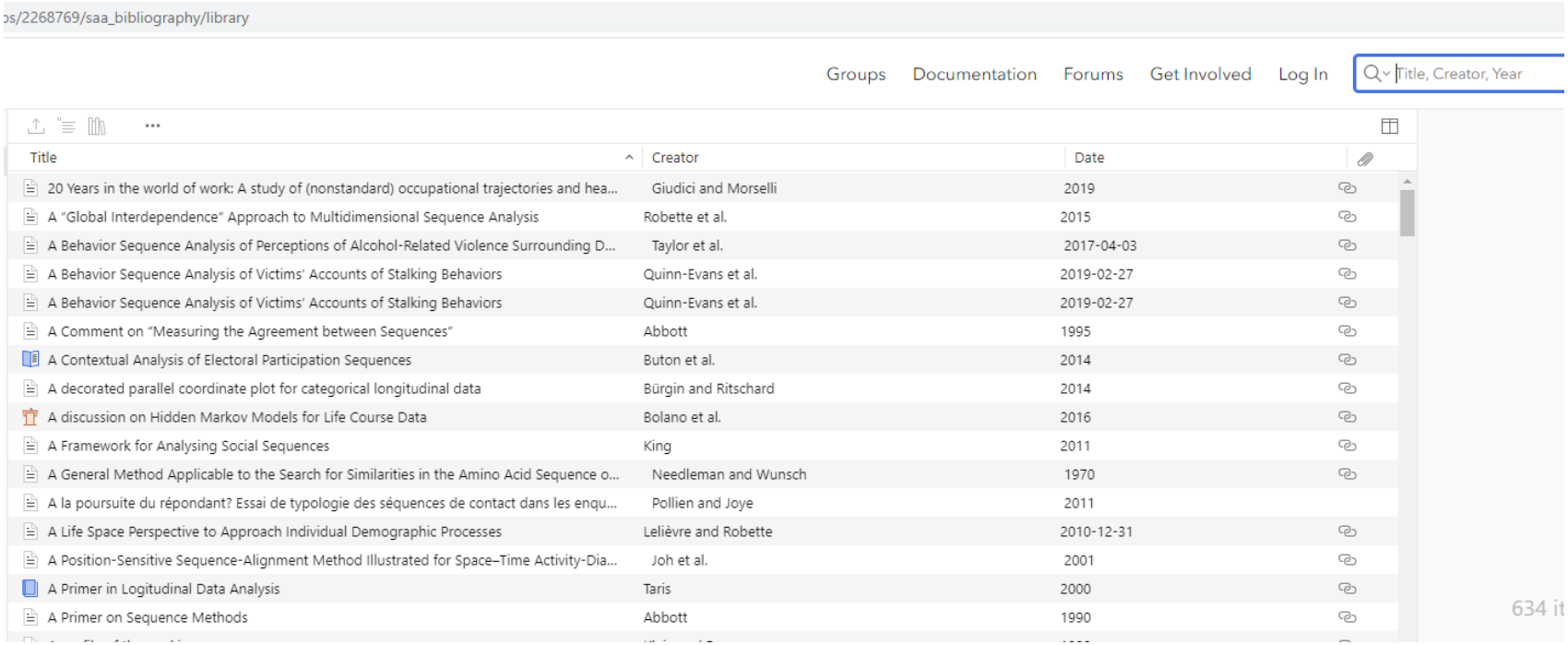
Despite some connections, the aims and methods of SA in social sciences strongly differ from those of sequence analysis in bioinformatics.

Contents [hide]

1 History

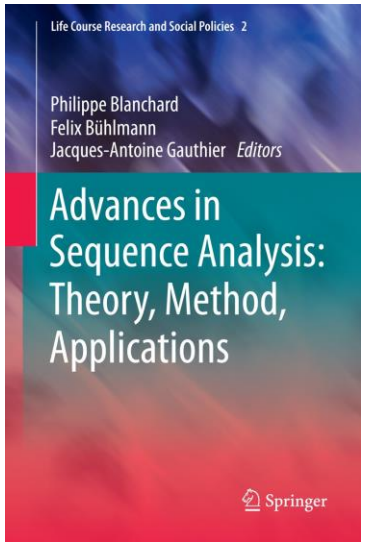
2 Domain-specific theoretical foundation

Sequence Analysis, where to find information



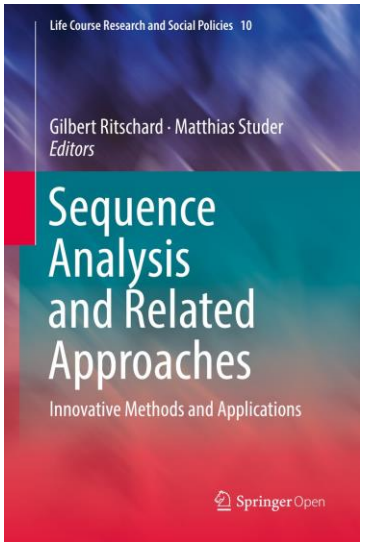
Title	Creator	Date
20 Years in the world of work: A study of (nonstandard) occupational trajectories and hea...	Giudici and Morselli	2019
A "Global Interdependence" Approach to Multidimensional Sequence Analysis	Robette et al.	2015
A Behavior Sequence Analysis of Perceptions of Alcohol-Related Violence Surrounding D...	Taylor et al.	2017-04-03
A Behavior Sequence Analysis of Victims' Accounts of Stalking Behaviors	Quinn-Evans et al.	2019-02-27
A Behavior Sequence Analysis of Victims' Accounts of Stalking Behaviors	Quinn-Evans et al.	2019-02-27
A Comment on "Measuring the Agreement between Sequences"	Abbott	1995
A Contextual Analysis of Electoral Participation Sequences	Buton et al.	2014
A decorated parallel coordinate plot for categorical longitudinal data	Bürgin and Ritschard	2014
A discussion on Hidden Markov Models for Life Course Data	Bolano et al.	2016
A Framework for Analysing Social Sequences	King	2011
A General Method Applicable to the Search for Similarities in the Amino Acid Sequence o...	Needleman and Wunsch	1970
A la poursuite du répondant? Essai de typologie des séquences de contact dans les enqu...	Pollien and Joye	2011
A Life Space Perspective to Approach Individual Demographic Processes	Lelièvre and Robette	2010-12-31
A Position-Sensitive Sequence-Alignment Method Illustrated for Space–Time Activity-Dia...	Joh et al.	2001
A Primer in Logitudinal Data Analysis	Taris	2000
A Primer on Sequence Methods	Abbott	1990

Sequence Analysis, where to find information



(2014)

(2015)

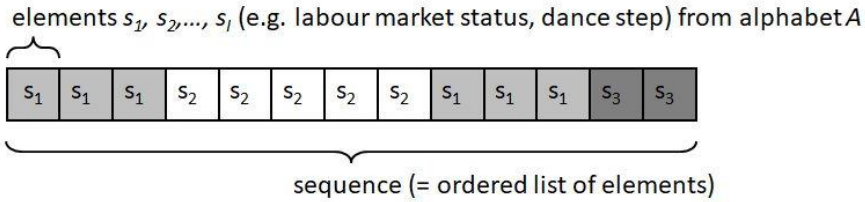


(2018)

(2022)

Sequence Analysis: Object of analysis

The object of analysis is a set of sequences, ordered lists of elements (s_1, s_2, …, s_l) taken from a finite alphabet (state space) A.

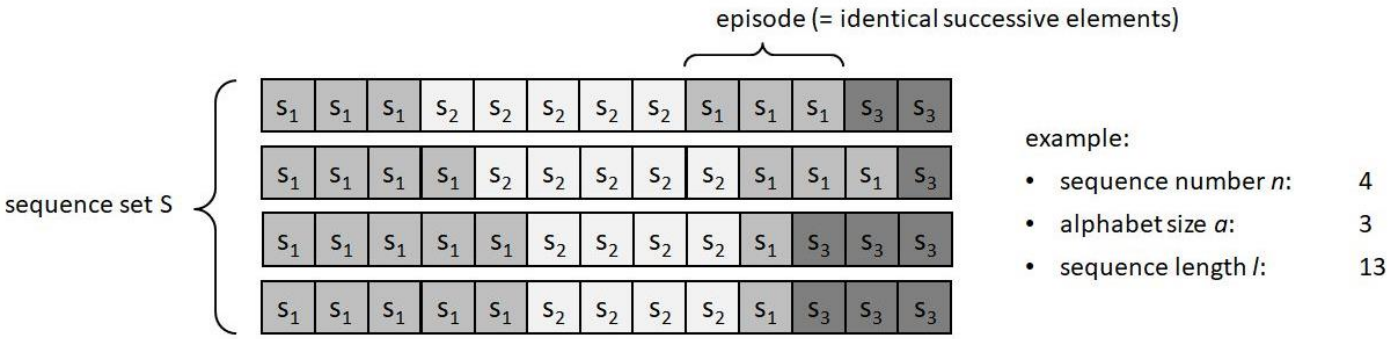


- Two requirements:
  - The elements of a sequence have to be ordered according to certain criteria (as we will see, in social sciences and demography usually a time dimension, such as age)
  - The elements (s_1, s_2, …, s_l) take value from a finite set A, called "alphabet" (e.g., in the case of labor market status, the elements could be "employed" and "unemployed").

Sequence Analysis: Object of analysis

A set of sequences n constitutes the object of analysis, our data frame. A few basic elements:

- n : number of sequences in the data frame
- a = |A| : size of the alphabet
- l : length of each sequence (It is possible to include sequences of different length)
- Transition : succession from a state i to a state j

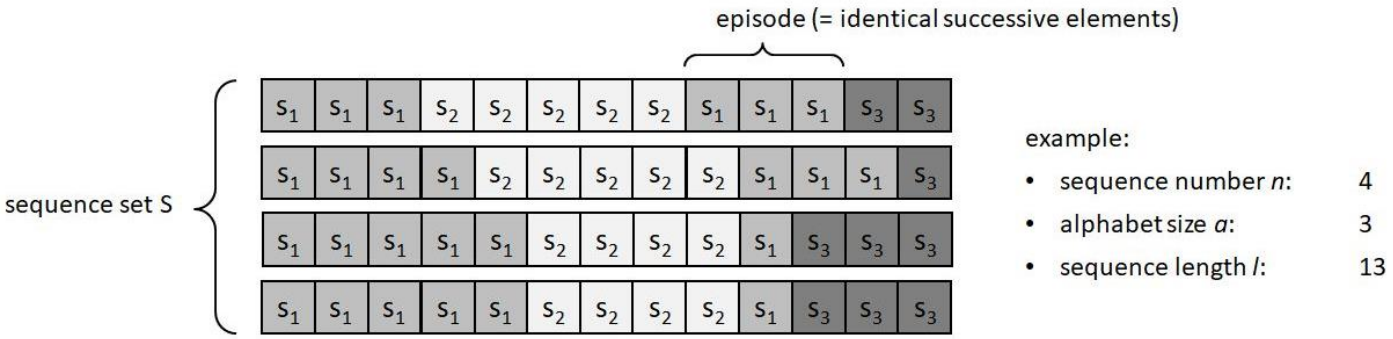


Source: https://en.wikipedia.org/wiki/Sequence_analysis_in_social_sciences

Sequence Analysis: Object of analysis

Sequence Analysis is the systematic and statistical analysis of a set of sequences. Traditionally, two main purposes:

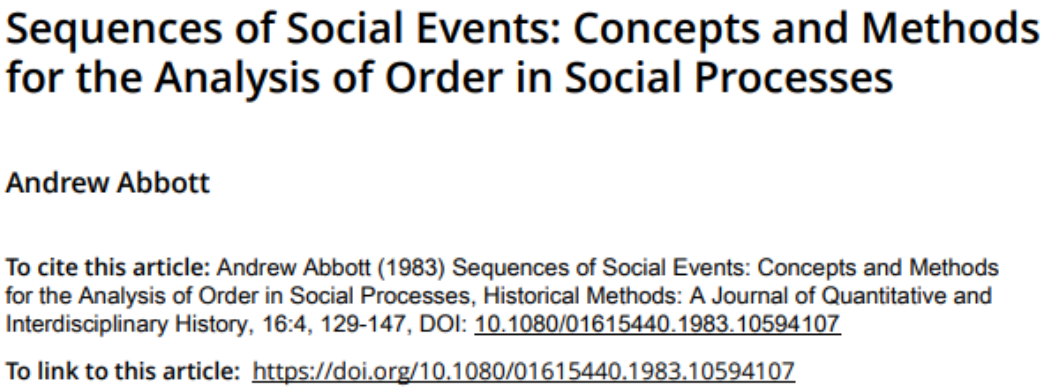
- Analysis of the sequences, computing indicators of individual sequences (e.g. number of transitions, diversity, complexity)
- Dissimilarity-based analysis: cluster analysis (Optimal Matching)



Source: https://en.wikipedia.org/wiki/Sequence_analysis_in_social_sciences

From where? A historical overview

- Sequence Analysis was introduced in the social sciences in the 1980s by Abbott, professor of sociology at the University of Chicago, from the biological sciences, where was developed to analyze DNA sequences, and information theory.



Sequences of Social Events: Concepts and Methods for the Analysis of Order in Social Processes

Andrew Abbott

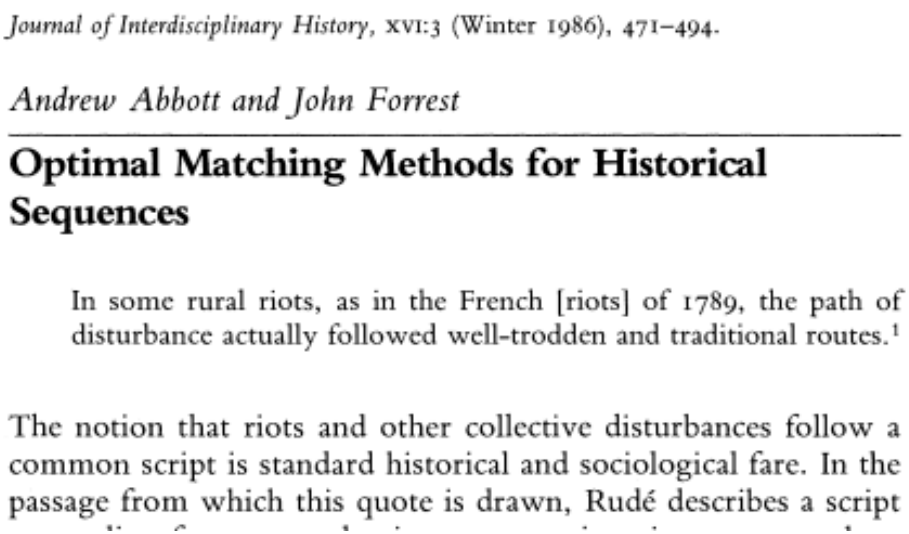
To cite this article: Andrew Abbott (1983) Sequences of Social Events: Concepts and Methods for the Analysis of Order in Social Processes, Historical Methods: A Journal of Quantitative and Interdisciplinary History, 16:4, 129-147, DOI: 10.1080/01615440.1983.10594107

To link to this article: https://doi.org/10.1080/01615440.1983.10594107

"How can we generalize about sequences of social events? In the past, such sequences have generally been the territory of historians. Yet, historians are little given to generalization. It is the sociologists who generalize, and they have said little about sequences. This paper aims to bridge the gap." (Abbott 1983, p. 129)

From where? A historical overview

- Introduction of Optimal Matching, which has become the most well-known method to compute distances between sequences in a clustering framework, as we will discuss later.



Journal of Interdisciplinary History, XVI:3 (Winter 1986), 471–494.

Andrew Abbott and John Forrest

Optimal Matching Methods for Historical Sequences

> In some rural riots, as in the French [riots] of 1789, the path of disturbance actually followed well-trodden and traditional routes.[1]

The notion that riots and other collective disturbances follow a common script is standard historical and sociological fare. In the passage from which this quote is drawn, Rudé describes a script

From where? A historical overview

In 1990, a seminal paper in AJS to introduce Optimal Matching in sociology:

Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers

Author(s): Andrew Abbott and Alexandra Hrycak

Source: *American Journal of Sociology*, Jul., 1990, Vol. 96, No. 1 (Jul., 1990), pp. 144-185

Published by: The University of Chicago Press

Stable URL: https://www.jstor.org/stable/2780695

This article introduces a method that measures resemblance between sequences using a simple metric based on the insertions, deletions, and substitutions required to transform one sequence into another. The method, called optimal matching, is widely used in natural science. The article reviews the literature on sequence analysis, then discusses the optimal matching algorithm in some detail. Applying this technique to a data set detailing careers of musicians active in Germany in the 18th century demonstrates the practical steps involved in the application of the technique and develops a set of typical careers that successfully categorize most of the actual careers studied by the authors.

From where? A historical overview

Abbott and Hrycak (1990)

- Analysis of careers and resemblance among career patterns:
  - Central in sociology: social mobility, organizational structure
  - Idea of analyzing careers as sequences: (i) individuals plan and structure their work histories, (ii) future career is not only influenced by the immediate present but also the actual sequence of experience in the past, and (iii) necessity to analyze work trajectories holistically
  - Are there common patterns of career trajectories? How are they produced?

From where? A historical overview

- Analysis of the careers of 595 musicians active in Germany during the Baroque and Classical eras, around 1660 – 1800.
  - Interesting description of the job system for musicians in 18th century Germany, a complex and unstable market stratified among towns and courts, where music was shaped largely by the local princes. Which career trajectories followed the major administrators (kapellmeisters and others) in town and courts?

A. Salieri:					
Main	50CRTCOM				
Amalgamated	16COPADM	36CRTKPM			
J. S. Bach:					
Main	1CRTVOC	5CHUORG	10CRTORG	6CRTKPM	27TWNMDR
Amalgamated	1CRTNON		4CRTKZM		27TWNCAN
					7COPKAP
					8TWNBAN
W. Mozart:					
Main	6CRTKZM	2OTHER	2CRTORG	5OTHER	4CRTCOM

CRTCOM: Court Composer
CRTVOC: Court Vocalist
CHUORG: Church organist
CRTORG: Court organist
CRTKPM: Court kapellmeister
TWNMDR: Town music director

Example: A successful career in court composition (Salieri), a variegated career leading to a prominent town position (Bach), and a short and relatively "unsuccessful" court career (Mozart).

From where? A historical overview

Still, we have to keep in mind that the scientific process is not linear. Longstanding tradition and foundations in sociological theories on regularities and sequencing of social phenomena.

For example:

- Durkheim: Norms and regularity of actions
- Giddens: Theory of structuration (social actors interface with each other in structured and sequenced patterns and routines)
- Bourdieu: Habitus, stable worldviews in guiding everyday action and producing predictable, orderly sequences of behavior.

From Salieri and Mozart to 3 decades of flourishing applications

Sociology
- Labor market
- Residential trajectories
- Time use
- Life course

Demography
- Transition to adulthood
- Partnership biographies
- Family formation
- Childbirth histories

Medicine
- Trajectories of chronic diseases

Political Sciences
- Pathways towards democratization
- Legislative processes
- National crises and revolutions

Geography
- Mobility studies
- Land use

Psychology
- Learning and interactions between individuals

Flourishing of Sequence Analysis, why?

Two primary reasons:

- Development of statistical software and data collection that made Sequence Analysis easy and feasible
  - Longitudinal and cross-sectional surveys that collected detailed individual-level data on careers, marital history, time use and time diary records.
  - Development of time-series data of larger social units, including organizations, neighborhoods, cities, and nations (e.g. World Values Survey, International Social Survey Programme)
  - Dedicated software: optimize, tda, SQ and SADI for Stata, TraMineR for R

Expansion of Sequence Analysis, why?

- Theoretical developments in sociology: the life course perspective.
  - Interdisciplinary program of study has been under development since the mid-1970s
  - Giele and Elder (1998) -> Four key elements: (i) Location in time and space (history and culture), (ii) Human agency (development of the individual), (iii) Linked lives (social relations), (iv) Timing (Intersection of age, period, and cohort)
  - Life course analysis as the statistical analysis of data on the timing of events (when do events happen?), their sequencing (in which order do events happen?), and their quantum (how many events happen?) (Billari 2002)

Expansion of Sequence Analysis, why?

Theoretical developments in sociology: the life course perspective.

- Life trajectories as the main object of analysis: family dynamics, transition to adulthood, and aging

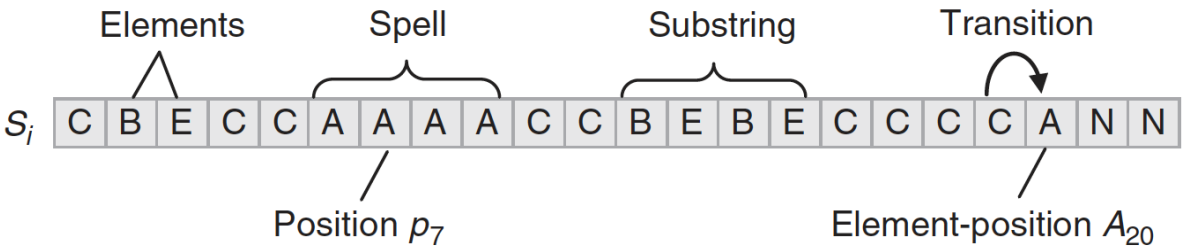
➢ Need for a holistic approach for life course analysis -> Sequence Analysis:

- Life courses as subject to accurate inter-temporal planning (internal calendars)
- Life courses as a holistic conceptual unit, contingent results of subsequent events.

Need for an algorithmic approach to describe and analyze the timing, sequencing, and quantum of life course events

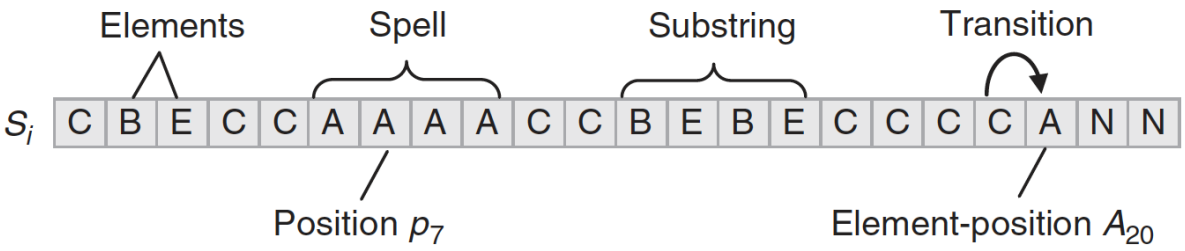
Formalizing the method: Basic concepts

The object of analysis is a set of sequences, ordered lists of elements (s_1, s_2, …, s_l) taken from a finite alphabet (state space) A.



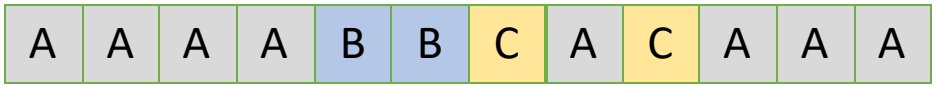
- **Position**: a key issue in sequence analysis is where and in what order elements appear in a sequence. Elements are ordered with respect to specific criteria, which can be e.g. temporal, spatial, hierarchical. Elements occupy a specific *position*.
- **Subsequence**: A set of ordered elements of a sequence, which need not be adjacent to each other in the parent sequence but have to appear in the same order. Example: CEAACCN.

Formalizing the method: Basic concepts

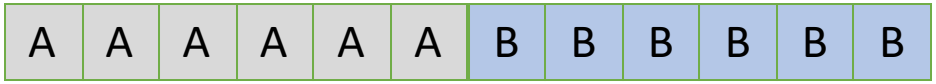


- **Substring**: A subsequence that is composed of only consecutive elements.
- **Spells:** A set of contiguous positions that all contain the same element. Spells are useful to report in a concise format a long sequence. For example, the sequence AAAACCNNNNB, can be reported as (A,4)-(C,2)-(N,4)-(B,1) (state-permanence-sequence format).

Describing the sequences: indicators



Sequence *x*



Sequence *y*

Alphabet A: {A,B,C}

a = |A|=3

l(x) = l(y) = 12

- **Number of transitions $l_d()$:** perhaps the simplest indicator, it is the number of state changes in the sequence.
- **Number of subsequences $\phi()$**: number of subsequences that can be extracted from the sequence.

Describing the sequences: indicators

A	A	A	A	B	B	C	A	C	A	A	A

Sequence *x*

A	A	A	A	A	A	B	B	B	B	B	B

Sequence *y*

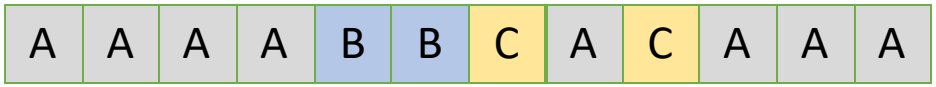
- **Shannon entropy:** The total time spent in each state characterizes the state distribution within a sequence. The entropy of this distribution can be seen as a measure of the diversity of its states. It can be computed as:

$$h(\pi_1, \dots, \pi_a) = -\sum_{i=1}^{a} \pi_i \ln \pi_i$$

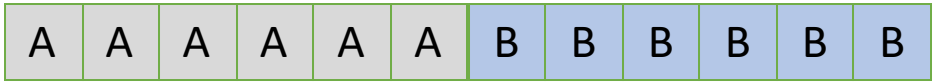
where a is the size of the alphabet and π_i is the proportion of the i^{th} state in the sequence. If the state remains the same during the whole sequence, the entropy equals 0, while the maximum entropy is reached when the same time is spent inside the sequence in each possible element of the alphabet.

e.g. Sequence x: - (0.66*ln0.66 + 0.16*ln0.16 + 0.16*ln0.16) = 0.86

Describing the sequences: indicators



Sequence *x*



Sequence *y*

- **Turbulence:** Composite measure that accounts for the number $\phi(x)$ of distinct subsequences and the variance $s_t^2(x)$ of the consecutive times t_j spent in the $l_d(x)$ distinct states. It can be computed as

$$T(x) = \log_2(\phi(x)\frac{s_{t,max}^2(x)+1}{s_t^2(x)+1})$$

where $s_{t,max}^2(x) = (l_d(x)-1)\big(1-\bar{t}(x)\big)^2$ and $\bar{t}(x)$ is the mean consecutive time spent in the distinct states.

For interpretation, sequences that have many distinct state and many state changes are more turbulent than sequences with fewer distinct states and/or fewer state changes.

Describing the sequences: indicators

A	A	A	A	B	B	C	A	C	A	A	A

Sequence *x*

A	A	A	A	A	A	B	B	B	B	B	B

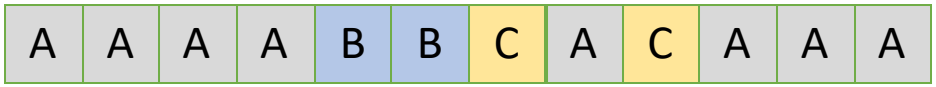
Sequence *y*

- **Complexity index:** Complexity index that combines the number of transitions in the sequence with the Shannon entropy. It can be compute as:

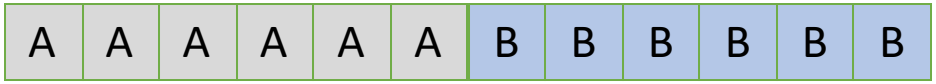
$$C(x) = \sqrt{\frac{(l_d(x) - 1)}{l(x) - 1}\frac{h(x)}{h_{max}}}$$

where $h_{max} = \ln(a)$, the theoretical maximum value of the entropy given the alphabet A.

Describing the sequences: overall statistics



Sequence *x*



Sequence *y*

Overall statistics computed on the overall set of sequences.

- **Mean time spent in each state:** a first synthetic information, the mean number of times each state that is observed in a sequence. It characterizes the overall state distribution.

e.g. mean time in the state A : (8+6)/2= 7

mean time in the state B : (2+6)/2= 4

mean time in the state C : (2+0)/2= 1

Describing the sequences: overall statistics

- **Transition rates:** transition rates between each couple of states (s_i, s_j), that is, the probability to switch at a given position from state s_i to state s_j. The transition rate $p(s_j|s_i)$ between states s_i and s_j can be computed as

$$p(s_j|s_i) = \frac{\sum_{t=1}^{L-1} n_{t,t+1}(s_i,\ s_j)}{\sum_{t=1}^{L-1} n_t(s_i)}$$

where L is the maximum observed sequence length, $n_t(s_i)$ is the number of sequences with state s_i at position t, and $n_{t,t+1}(s_i,\ s_j)$ the number of sequences with state s_i at position t and with state s_j at position $t+1$

Dissimilarity-based analysis: cluster analysis

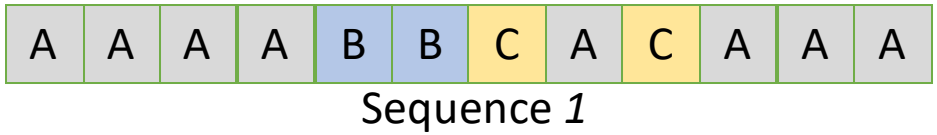
Central in sequence analysis since the first developments. Abbott and Tsay (2000) describe Sequence Analysis essentially as a three-step process:

(i) Coding individual observations as sequences of states

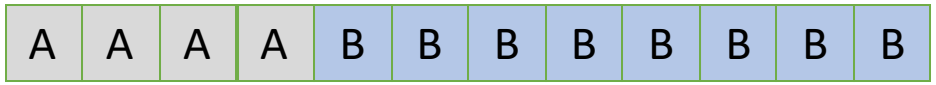
(ii) Measuring pairwise dissimilarities between sequences (using Optimal Matching)

(iii) Clustering the sequences from pairwise dissimilarities to find typical patterns (e.g. typical patterns of revolutions, typical patterns of musician career in Baroque Germany)

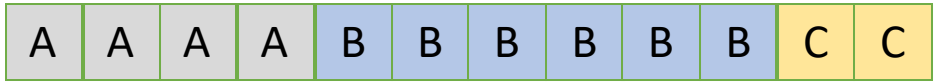
Measuring pairwise dissimilarities



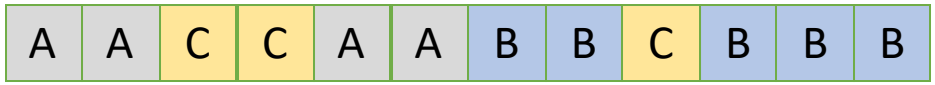
Sequence *1*



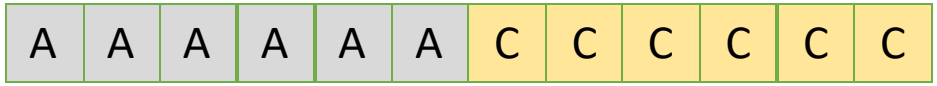
Sequence *2*



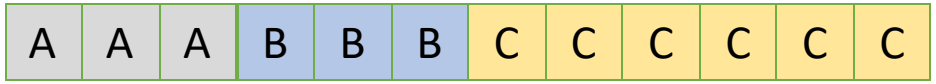
Sequence *3*



Sequence *4*



Sequence *5*



Sequence *6*

Measuring pairwise dissimilarities

Optimal Matching:

An algorithm to compute the distance between two sequences as the minimal "effort" of transforming one sequence into the other. To do so, we have to assign a penalty, a "cost", for each operation or transformation we need to employ. The sum of these costs provide a measure of distance between sequences.

In classical Optimal Matching, the possible operations are element deletion, insertion, and substitution. We should then define the cost associated to each of these operations – let's take an example by setting deletion, insertion, and substitution cost all equal 1.

S_i^*	A	A	A	A	B	B	B	B	
S_j^*	A	*B*	*B*	*C*	*C*	*D*	*E*	*E*	
Cost:	0	1	1	1	1	1	1	1	= 7

Measuring pairwise dissimilarities

Optimal Matching:

S_i	A	A	A	A	B	B	B	B				
S_j	ϕ	ϕ	ϕ	A	B	B	*C*	*C*	~~D~~	~~E~~	~~E~~	
Cost:	1	1	1	0	0	0	1	1	1	1	1	= 8

For S_j, the letter ϕ is used to signify places where insertions occurred, the strikethroughs (over the D and the Es) represent deletions, and the underlined and italicized elements (the Cs) signify elements that were substituted out. We can treat this sum of 8 as a measure of the distance between S_i and S_j .

Measuring pairwise dissimilarities

Optimal Matching, how to define costs?

In the literature, huge debate on how costs should be quantified.

Basic approaches:

- Levenshtein Distance: same costs, 1, for both indels and substitutions
- Levenshtein II Distance: cost 1 for insertions and deletions and cost 2 for substitutions

Recommended approach:

- Cost 1 for insertions and deletions and substitution cost inversely proportional to the transition rate probabilities $(2 - p(s_i|s_j) - p(s_j|s_i))$.

Cluster analysis

After defining a cost metric, it is possible to compute the dissimilarity matrix D, which contains the distances between subjects' sequences (several algorithms available to run Optimal Matching).

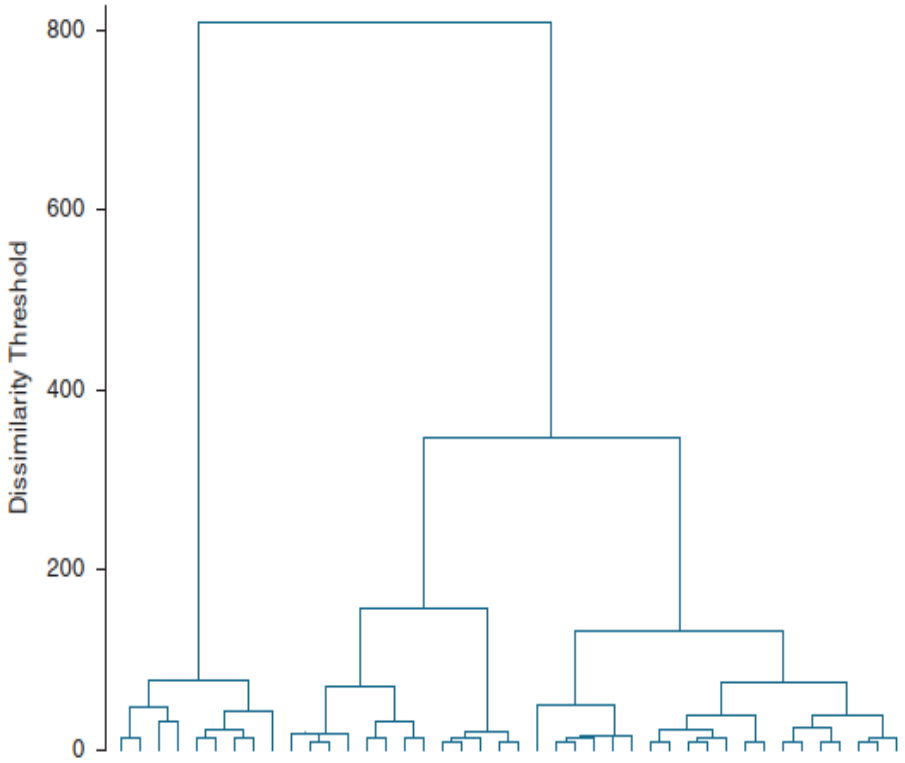
Then, we can use the dissimilarity matrix in a clustering algorithm to compute groups of similar sequences, sets of cases that are less distant from (more similar to) each other. The assumption is that these different groups in turn reflect distinct classes of sequences.

Again, there is no consensus and numerous approaches are available. However, agglomerative hierarchical clustering approaches become widely used in the Sequence Analysis community.

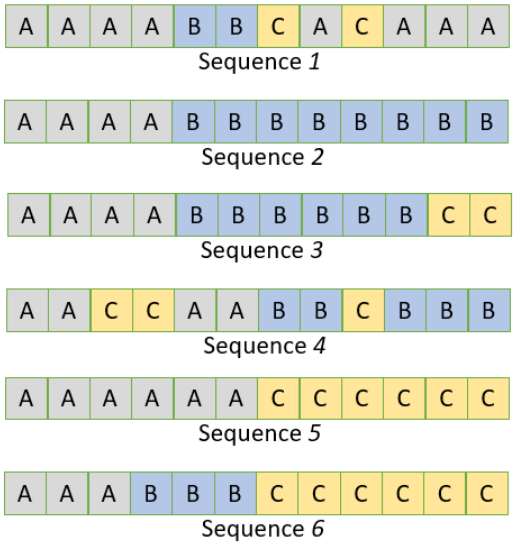
Cluster analysis

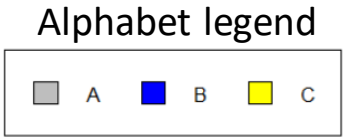
Agglomerative hierarchical clustering:

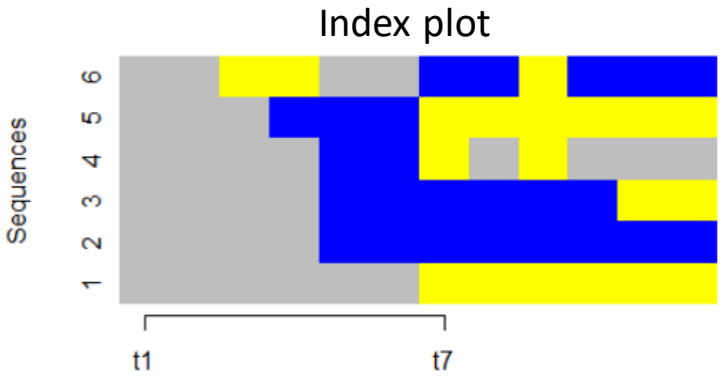
- Initially consider all cases are singular clusters
- Iteratively combine groups that are “close” to each other to some specified criteria until we eventually end up at a single large cluster.
- Regardless of which linkage criteria are used, hierarchical clustering leads to a set of nested clusters that can be represented in a tree diagram, a dendrogram.

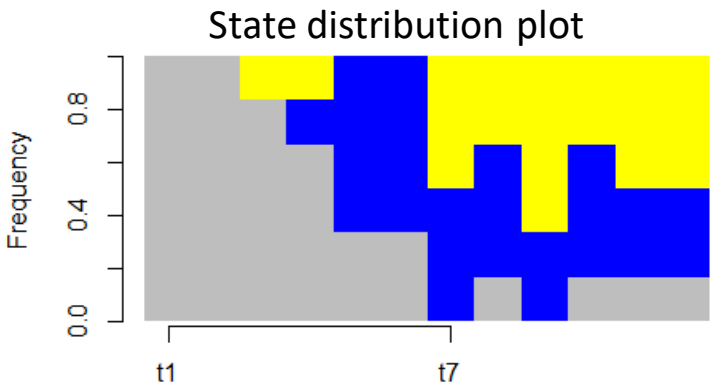


Sequence Analysis: data visualization



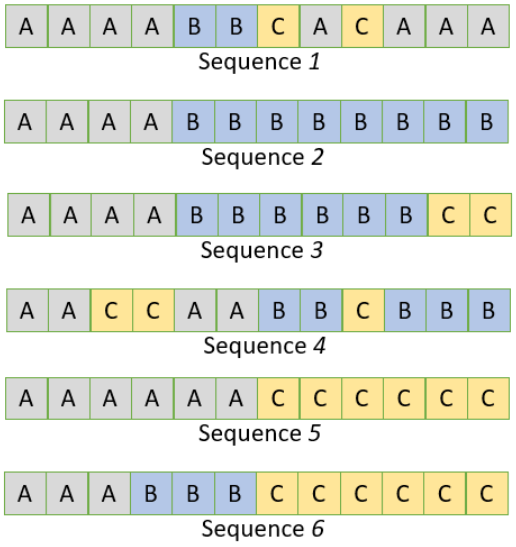




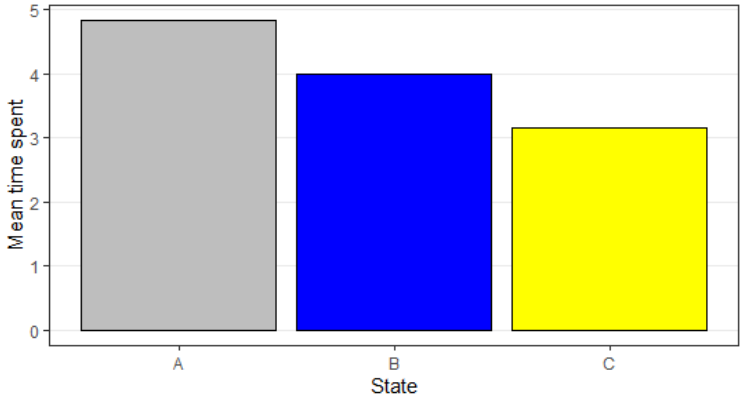



```
> seqstatd(data.seq)
  [State frequencies]
   t1 t2   t3   t4   t5   t6  t7   t8   t9  t10  t11  t12
A   1  1 0.83 0.67 0.33 0.33 0.0 0.17 0.00 0.17 0.17 0.17
B   0  0 0.00 0.17 0.67 0.67 0.5 0.50 0.33 0.50 0.33 0.33
C   0  0 0.17 0.17 0.00 0.00 0.5 0.33 0.67 0.33 0.50 0.50
```

Sequence Analysis: data visualization



Mean time spent in ache state



Transition rate matrix

```
> round(tr, 4)
       [-> A] [-> B] [-> C]
[A ->] 0.7143 0.1786 0.1071
[B ->] 0.0000 0.8182 0.1818
[C ->] 0.1875 0.0625 0.7500
```

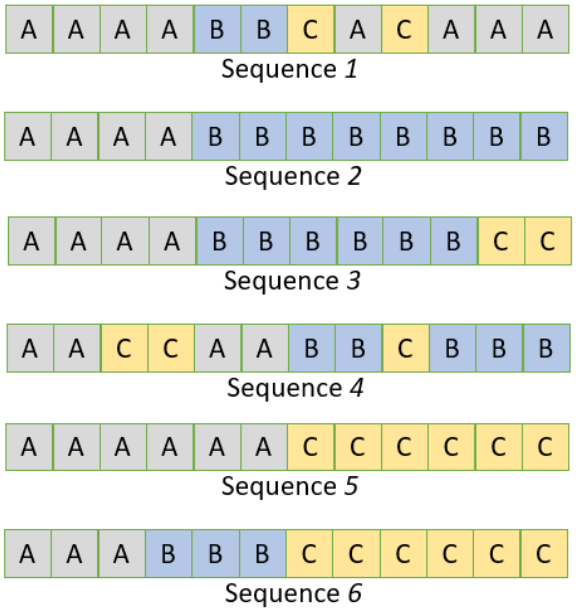
Optimal Matching distance

```
> dist.om1 <- seqdist(data.seq, method = "OM", indel = 1, sm = submat)
 [>] 6 sequences with 3 distinct states
 [>] checking 'sm' (size and triangle inequality)
 [>] 6 distinct  sequences
 [>] min/max sequence lengths: 12/12
 [>] computing distances using the OM metric
 [>] elapsed time: 0 secs
> dist.om1
          1         2         3         4         5         6
1  0.000000 10.797078  8.000000  9.398539 10.464286  8.642857
2 10.797078  0.000000  3.511364  5.755682 14.176948 10.778409
3  8.000000  3.511364  0.000000  5.755682 10.665584  7.267045
4  9.398539  5.755682  5.755682  0.000000 12.189123 11.088474
5 10.464286 14.176948 10.665584 12.189123  0.000000  5.464286
6  8.642857 10.778409  7.267045 11.088474  5.464286  0.000000
```

LAB SESSION

Lab Session 1

An easy toy example using our set of 6 sequences



Lab Session 2.
Stratified pathways to Italy's "Latest-Late" Transition to Adulthood

During the last decades, society as a whole has undergone substantial social, economic, and demographic changes.

The transition to adulthood progressively moved from an **early**, **contracted**, and **simple** pattern - dominant in the 1950s and 1960s - to a **late**, **protracted**, and **complex** pattern - dominant in contemporary developed countries.

Modernization -> **Individualization** and **de-standardization** of life course trajectories.

- The **Second Demographic Transition**, a global and "individual choice" paradigm
- Growing perspectives: centrality of **social stratification**, the **gender revolution**, and **context**ual opportunities and constraints.

The context: Italy in a comparative perspective

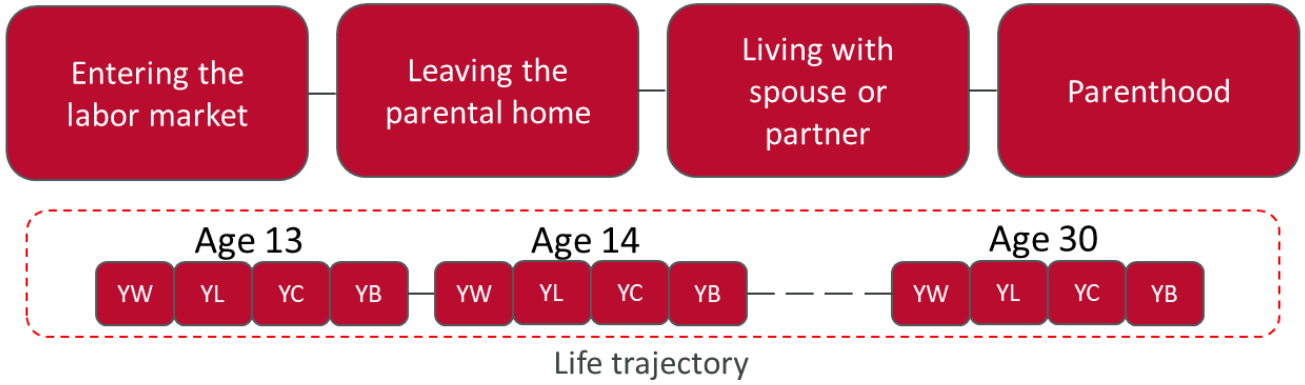
	Italy	Northern Europe	Baltic States	Western Europe	Eastern Europe	Southern Europe
Entering the job market (Age 22)						
Men	59.0%	77.8%	89.2%	74.5%	71.3%	69.8%
Women	41.6%	72.8%	72.4%	70.8%	61.2%	57.6%
Leaving the parental home (Age 22)						
Men	29.8%	86.6%	74.0%	61.3%	37.0%	37.8%
Women	28.9%	90.9%	84.1%	77.2%	47.6%	50.9%
Living with a partner of a spouse (Age 30)						
Men	54.0%	82.9%	81.5%	83.0%	71.0%	65.5%
Women	57.3%	89.6%	91.6%	89.9%	81.7%	78.5%
Parenthood (Age 30)						
Men	19.8%	37.3%	56.2%	37.1%	49.9%	24.4%
Women	34.3%	56.2%	73.8%	59.7%	66.2%	44.1%

Note: **Cohort of respondents born between 1978 and 1988**. Northern Europe: Denmark, Finland, Norway, Sweden, Iceland. Baltic States: Estonia, Lithuania, Latvia. Western Europe: Austria, Belgium, United Kingdom, France, Germany, Ireland, The Netherlands, Switzerland. Southern Europe: Portugal, Spain.
Source: Authors' elaborations of the European Social Survey Round 9, 2018. Post-stratification weights have been applied for between-countries analyses.

- Italy, a **unicum** in the European panorama?
- The **"latest-late"** transition to adulthood associated to the "lowest-low" fertility rates.

Data and methods

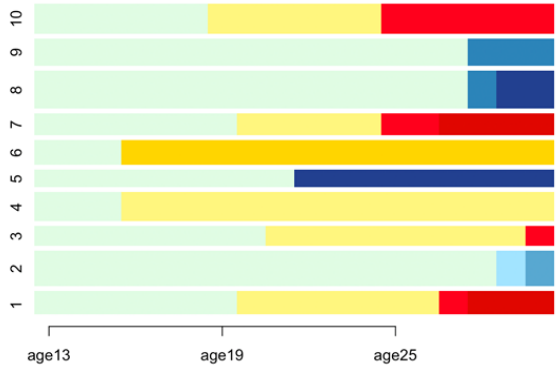
- European Social Survey 2018-19, "Timing of Life" rotating module



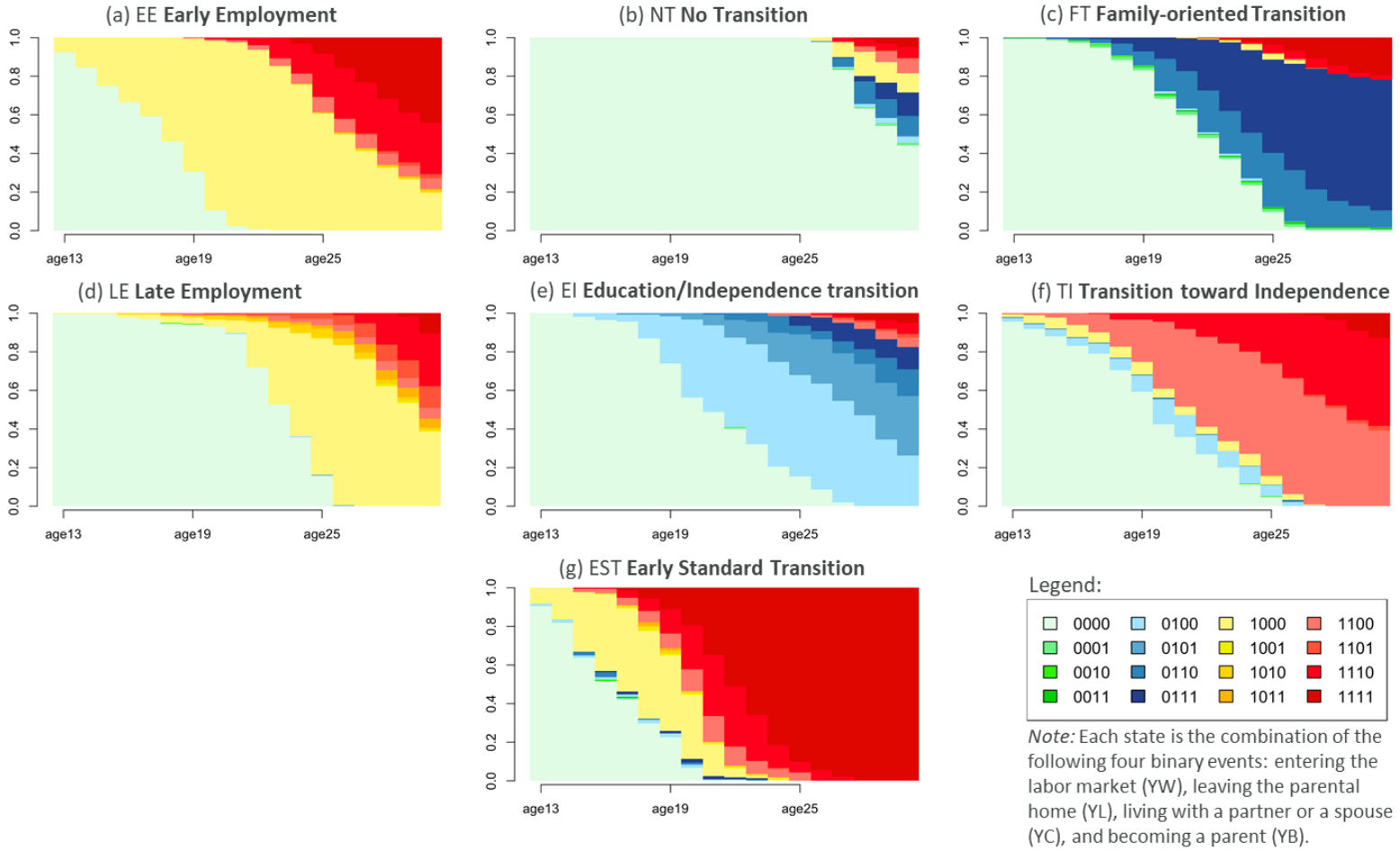
Legend:

0000	0100	1000	1100
0001	0101	1001	1101
0010	0110	1010	1110
0011	0111	1011	1111

Note: Each state is the combination of the following four binary events: entering the labor market (YW), leaving the parental home (YL), living with a partner or a spouse (YC), and becoming a parent (YB).



Sequence Analysis



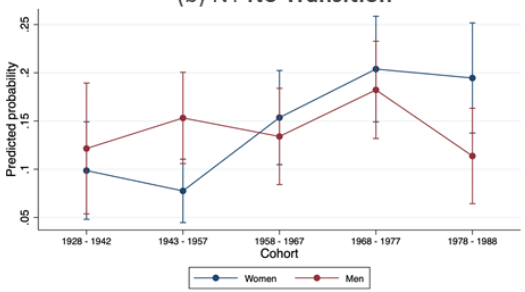
Note: Each state is the combination of the following four binary events: entering the labor market (YW), leaving the parental home (YL), living with a partner or a spouse (YC), and becoming a parent (YB).

Multinomial logistic regression

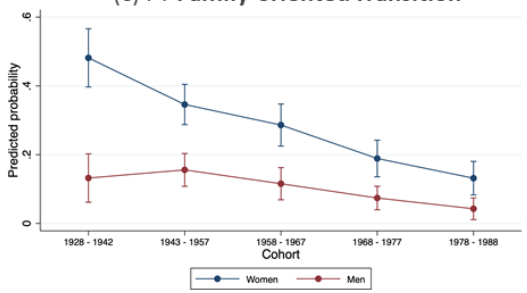
Predicted probabilities by gender and cohort

(a) EE **Early Employment**

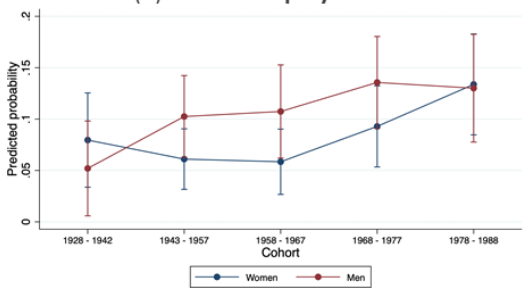
(b) NT **No Transition**



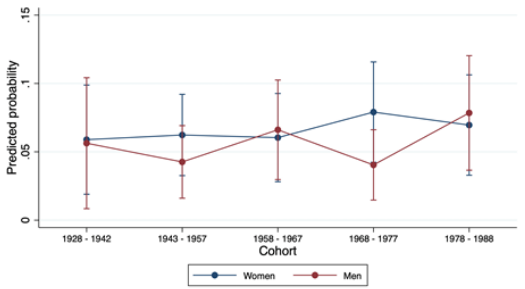
(c) FT **Family-oriented Transition**



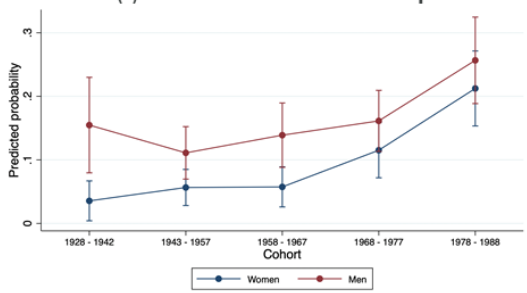
(d) LE **Late Employment**



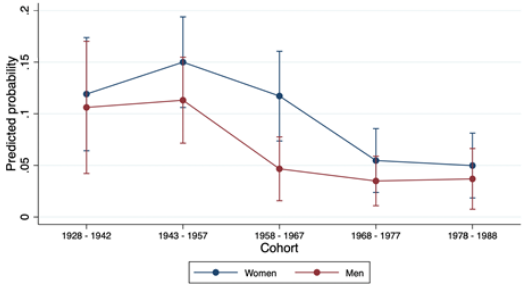
(e) EI **Education/Independence transition**



(f) TI **Transition toward Independence**



(g) EST **Early Standard Transition**



Multinomial logistic regression

Average marginal effects by gender, cohort, and education of parents

Variable	EE Early Employment	EI Education/ Independence	EST Early Standard	FT Family-oriented	LE Late Employment	NT No Transition	TI Independence
			Men				
Cohort	-	-	-	-	-	-	-
(non reported here for sake of clarity, available upon request)							
Education of parents *(low as reference)*							
Medium-low	-0.008	-0.030*	-0.019	-0.024	0.076***	0.041	-0.037
Medium-high	-0.080	-0.025	-0.041*	-0.036	0.078**	0.117**	-0.012
High	-0.165***	-0.008	0.023	-0.068**	0.070	0.120*	0.028
			Women				
Cohort	-	-	-	-	-	-	-
(non reported here for sake of clarity, available upon request)							
Education of parents *(low as reference)*							
Medium-low	0.060*	-0.007	-0.038	-0.138***	0.053**	0.026	0.044**
Medium-high	-0.035	0.012	-0.096***	-0.119***	0.108***	0.046	0.084**
High	-0.116**	0.038	-0.067**	-0.150***	0.076*	0.100*	0.119**

*p<0.05, **p<0.01, ***p<0.001

Source: Authors' elaborations of the European Social Survey Round 9, 2018. Post-stratification weights have been applied for country-level analysis.

R Code