Categorical sequence analysis with optimal matching: an application to the relationships history of women over 40

Adriana Clavijo Daza

Statistics and Data Science Master's, Universität Bern

2022-06-02

Motivation

Understand the similarities and differences in the romantic relationships history of a group of women over 40 years old and use this information as a predictor of personality traits.

Women 40+ Healthy Aging Study (i)

Universität Zürich[UZH]

Dynamics of Healthy Aging

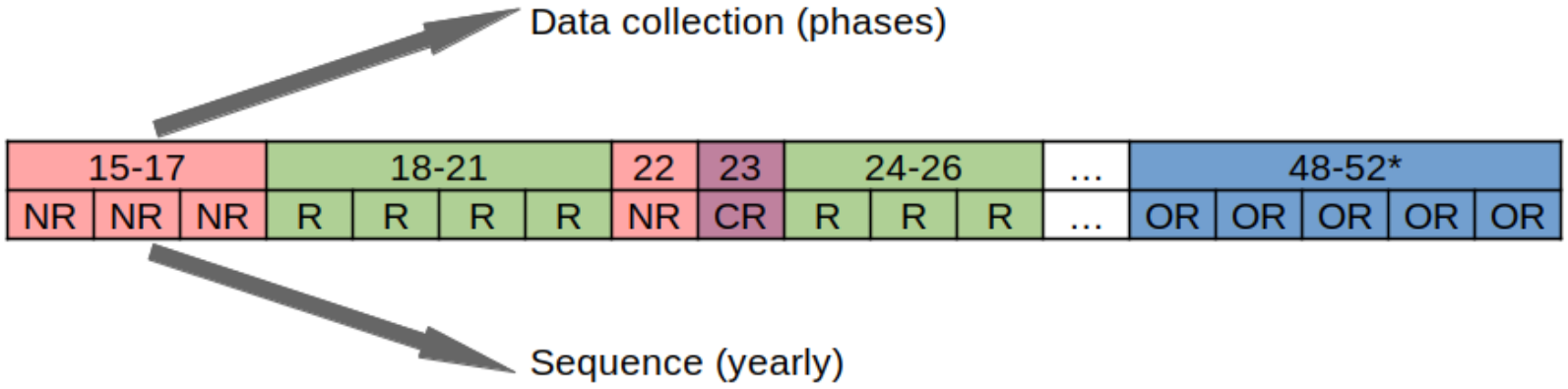
- Data from 250 individuals collected between June 2017 and February 2018.
- Psychometric instrument to obtain information about the history of romantic relationships of women aged between 40 and 75 years.

Women 40+ Healthy Aging Study (ii)

- Information about relationship phases starting from the age of 15 years until the current age at the time of the data collection.
- The phases are defined by the start and end age and for each phase and information about civil status, relationship status, living situation, children and quality of the relationship was collected.
- Additional information about the individuals was collected, in particular, scores for personality scores.

Data example

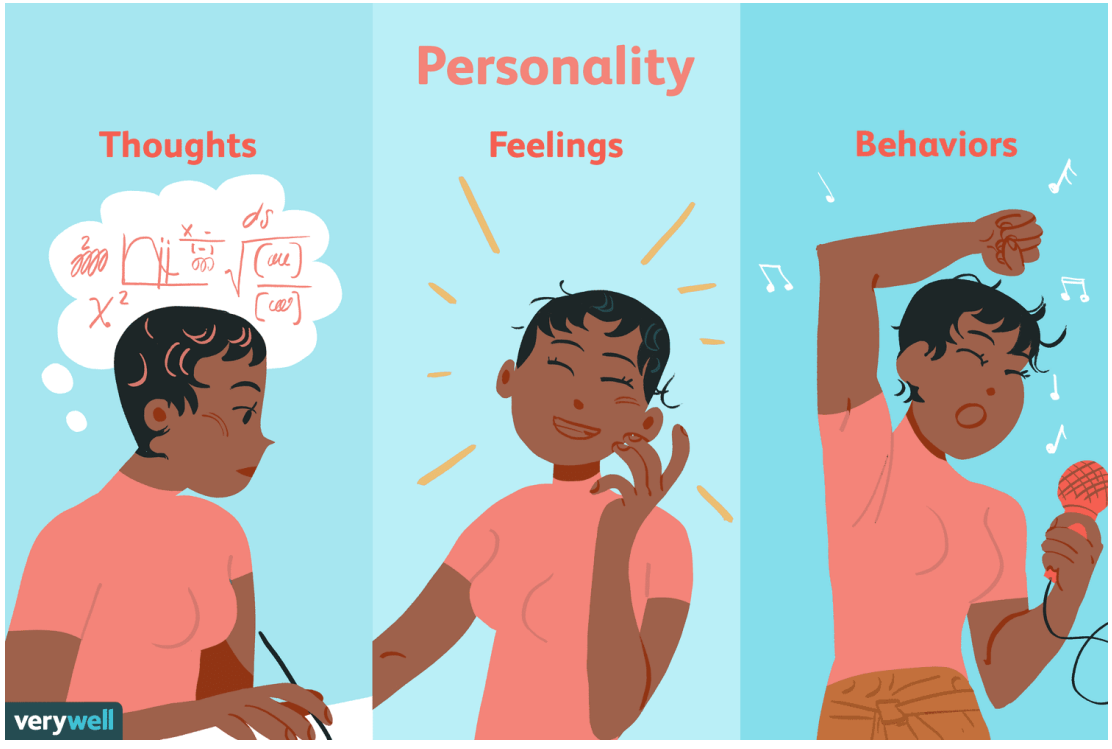
Consider the relationship status:



*Current age

- No relationship (NR)
- In a relationship (R)
- Open relationship (OR)
- Changing relationships (CR)

What is personality?



Personality refers to the enduring characteristics and behavior that comprise a person's unique adjustment to life, including major traits, interests, drives, values, self-concept, abilities, and emotional patterns.

The “Big Five” personality traits

O	**Openness to Experience**	Appreciation for art, new ideas, variety of experiences imagination and curiosity	„I have many different interests“
C	**Conscientiousness**	Tendency towards self-discipline and striving for achievement against measures or outside expectations.	„I always follow my plans“
E	**Extraversion**	Gain energy from external situations and means, enjoy a breadth of activities and assert their viewpoints	„I am more the quite type“ (reverse coded)
A	**Agreeableness**	Value social harmony and getting along with others, optimistic, kind and generous towards others	„I am cooperative and prefer working in teams over competition“
N	**Neuroticism**	Tendency to experience negative emotions, such as anger, anxiety, or depression. Low tolerance of stress	„I worry a lot“

Research question

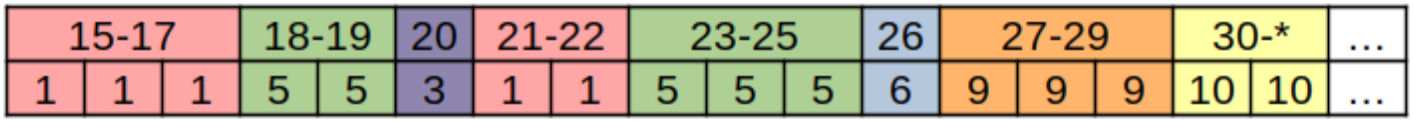
- Can we get a good prediction of personality scores based on the relationship history sequences?

Data pre-processing

- Manual corrections of several inconsistent and incomplete records.
- Several additional automatic checks to identify sequences with inconsistent data.
- Corrections based on secondary data source.
- Identification and selection of the variables and patterns that provide a wider perspective of the individuals' situations.
- In total, 239 individuals are considered for the analysis.

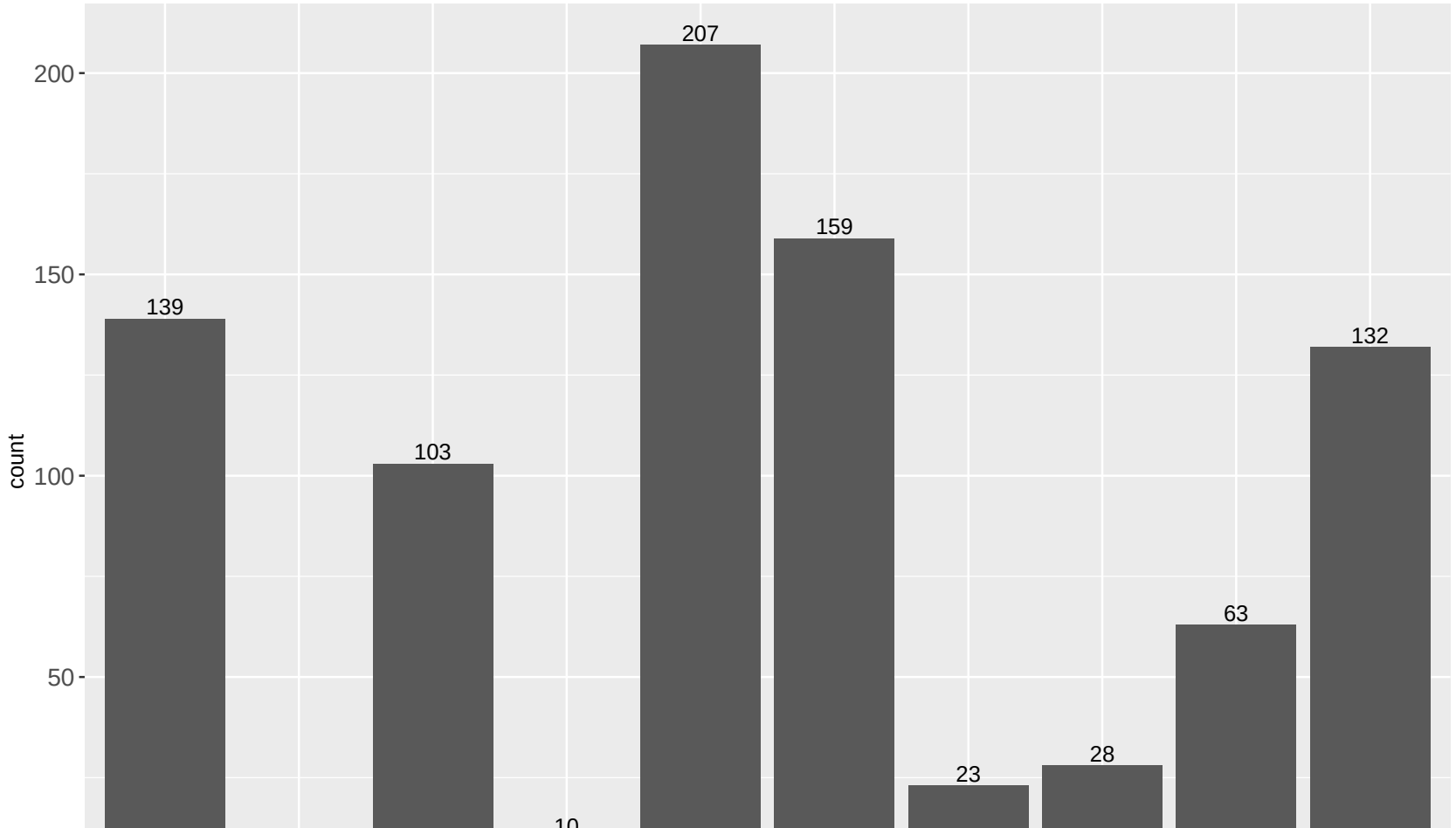
Considered states

- 1 = Single + no children
- 2 = Single + children
- 3 = Changing relationships + no children
- 4 = Changing rel. + children
- 5 = Relationship + living apart + no children
- 6 = Relationship + living together + no children
- 7 = Relationship + living apart + children
- 8 = Relationship + living together + children
- 9 = Married + no children
- 10 = Married + children



Distribution of states

```
## Warning: The dot-dot notation (`..count..`) was deprecat
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where
## generated.
```



Optimal Matching (OM)

- Technique used in social sciences for the comparison of sequences of categorical states indexed by time.
- Applications on life course and career path analysis.
- Uses the Needleman-Wunsch algorithm, that was developed to compare biological sequences.
- The Needleman-Wunsch algorithm is an application of dynamic programming, an iterative method that simplifies an optimization problem by breaking it into a recursion of smaller problems.

Example (i)

Analyzing Sequence Data: Optimal Matching in Management Research (T. Biemann and D. K. Datta)

- Goal: study career paths of deans at US business schools.
- Data source: 149 CVs of deans including public and private business schools.
- Coded into yearly data with the states: administration (A), corporation (C), faculty (F), government (G).

Table 2. Examples of Career Paths of U.S. Business School Deans.

	Career Path
Dean 1	F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-A-A-A-A-A-A-A-A-A-A-A-A-A-A-A-A-A-A
Dean 2	F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-A-A
Dean 3	C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C
Dean 4	F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-A-F-F-C-C-F-F-F-F-G-G-G-G-G-G-G-G-A-A-A
Dean 5	C-C-C-C-C-C-C-C-C-F-F-F-F-A-A-A-A-A-A-A-A-A-A-F-F-F-F-F-F-F-A-A

Example (ii)

Cost matrix:

Table 3. Absolute Frequency, Relative Frequency, and Substitution Costs Between States.

			Substitution Costs				
	Absolute Frequency	Relative Frequency (%)	F	A	C	G	NA
F	2,454	54.5	0.000	1.891	1.893	1.870	2.000
A	1,144	25.4	1.891	0.000	1.977	1.971	2.000
C	693	15.4	1.893	1.977	0.000	1.939	2.000
G	200	4.4	1.870	1.971	1.939	0.000	2.000
NA	14	0.3	2.000	2.000	2.000	2.000	0.000
Sum	4,505	100.00	(indel costs = 1)				

Example (iii)

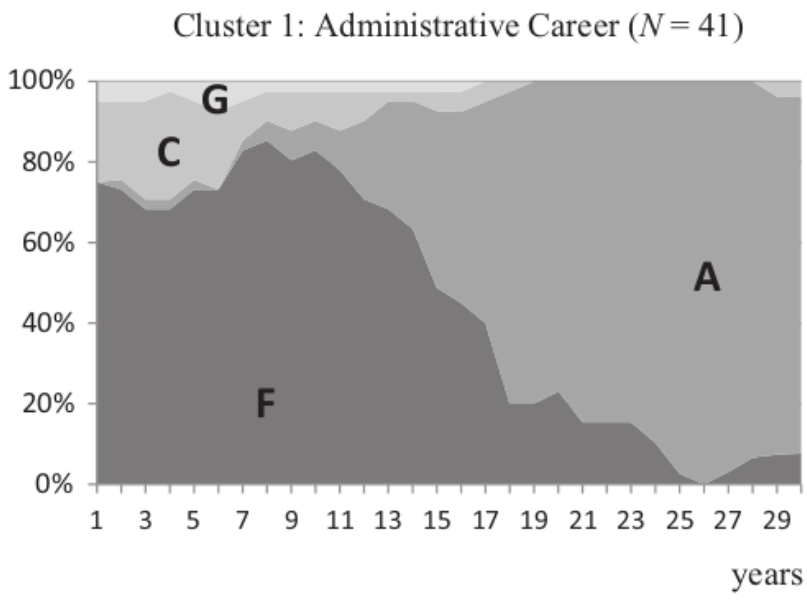
Distance/dissimilarities matrix for five deans:

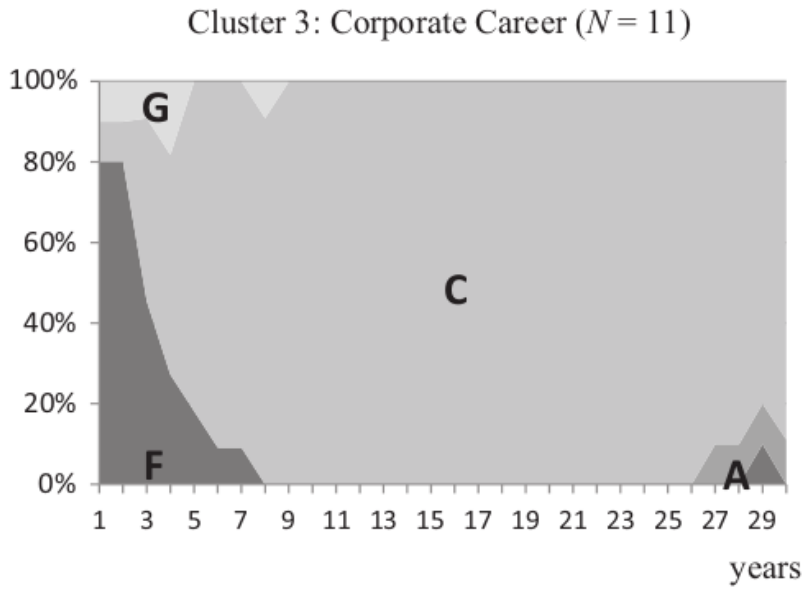
Table 4. Distance Matrix for Five Deans.

	Dean 1	Dean 2	Dean 3	Dean 4	Dean 5
Dean 1	—				
Dean 2	23.13	—			
Dean 3	69.60	61.07	—		
Dean 4	28.52	18.22	64.14	—	
Dean 5	35.38	33.27	47.55	40.67	—

Example (iv)

Two of the five resulting clusters:





The OM algortihm (i)

- Set of n states: $S = \{s_1, \ldots, s_n\}$
- Sequence of size $t > 0$: $X = (x_1, \ldots, x_t)$, with $x_i \in S$ for $i = 1, \ldots, t$.
- **S** is the set of all possible sequences with states belonging to S.

Objective: Find the optimal way to align these two sequences

- Let $X, Y \in \mathbf{S}$ be two sequences of size t_1 and t_2, respectively.
- Define an empty array F of size $(t_1 + 1) \times (t_2 + 1)$

The OM algortihm (ii)

1: $F(1,1) \leftarrow 0$
2: **for** $j \leftarrow 2, t_2 + 1$ **do**
3: $\quad F(1,j) \leftarrow F(1,j-1) + d$
4: **end for**
5: **for** $i \leftarrow 2, t_1 + 1$ **do**
6: $\quad F(i,1) \leftarrow F(i-1,1) + d$
7: **end for**
8: **for** $i \leftarrow 2, t_1 + 1$ **do**
9: $\quad$ **for** $j \leftarrow 2, t_2 + 1$ **do**
10: $\quad\quad F(i,j) \leftarrow$
$\min\{F(i-1,j)+d, F(i,j-1)+d, F(i-1,j-1)+k(y_{i-1}, x_{j-1})\}$
11: $\quad$ **end for**
12: **end for**

The OM algortihm (iii)

- d is the cost of inserting a gap (indel cost).
- $k(y_{i-1}, x_{j-1})$ is the cost associated to change from the state y_{i-1} to x_{j-1}.
- These costs are defined in a matrix K of size $n \times n \rightarrow$ cost matrix.
- Lines 1-7 of the algorithm correspond to initialization.
- The remaining lines of the algorithm correspond to the row-wise recursion to fill the array F.
- When F is completely filled, the value $F(t_1 + 1, t_2 + 1)$ corresponds to the optimal cost of aligning the sequences X and Y.

Cost matrix (i)

The `R` package `TraMineR` provides several functions to work with sets of sequences. The package implements OM and offers several methods for computing the cost matrix K.

Cost matrix (ii)

- Transition rates (`TRATE`):

The substitution cost between states s_i and s_j, $1 \leq i, j \leq n$, is calculated as:

$$K(s_i, s_j) = c - P(s_i|s_j) - P(s_j|s_i), \tag{1}$$

where $P(s_i|s_j)$ is the probability of transition from state s_j in time t to s_i in time $t + 1$ and c is a constant, set to a value such that $0 \leq K(s_i, s_j) \leq 2$).

Cost matrix (iii)

- Chi-squared distance (`FUTURE`):

$$K(s_i, s_j) = ChiDist(\mathbf{P_i}, \mathbf{P_j}), \tag{2}$$

where $\mathbf{P}_\cdot = (P(s_1|.), \ldots, P(s_n|.))'$

Cost matrix (iv)

- Relative frequencies (`INDELS` and `INDELSLOG`):

$$K(s_i, s_j) = indel_i + indel_j, \tag{3}$$

where $indel_i = 1/f_i$ for method `INDEL`, $indel_i = \log[2/(1 + f_i)]$ and f_i is the relative frequency of the state s_i for $i = 1, \ldots, n$.

Example (i)

- $S =$ the alphabet
- $X = \{S, E, N, D\}, Y = \{A, N, D\} \in \mathbf{S}$
- $d = 2$

$$K(i,j) = \begin{cases} 0 & \text{if } i = j, \\ 3 & \text{otherwise} \end{cases}$$

Example (ii)

		S	E	N	D
	0	2	4	6	8
A	2				
N	4				
D	6				

- $F(2,2) = min\{F(1,2) + d, F(2,1) + d, F(1,1) + k(y_1, x_1)\} = min\{2+2, 2+2, 0+3\} = 3$
- $F(2,3) = min\{F(1,3) + d, F(2,2) + d, F(1,2) + k(y_1, x_2)\} = min\{4+2, 3+2, 2+3\} = 5$
- $F(2,4) = min\{F(1,4) + d, F(2,3) + d, F(1,3) + k(y_1, x_3)\} = min\{6+2, 5+2, 4+3\} = 7$
- $F(2,5) = min\{F(1,5) + d, F(2,4) + d, F(1,4) + k(y_1, x_4)\} = min\{8+2, 7+2, 6+3\} = 9$

Example (iii)

		S	E	N	D
	0	2	4	6	8
A	2	3	5	7	9
N	4				
D	6				

- $F(3,2) = min\{F(2,2) + d, F(3,1) + d, F(2,1) + k(y_2, x_1)\} = min\{3 + 2, 4 + 2, 2 + 3\} = 5$
- $F(3,3) = min\{F(2,3) + d, F(3,2) + d, F(2,2) + k(y_2, x_2)\} = min\{5 + 2, 5 + 2, 3 + 3\} = 6$
- ...

Cost matrix (ii)

Status	Single+no ch.	Single+ch.	Changing rel.+no ch.	Changing rel.+ch.	Rel. +apart+no ch.	Rel. +together +no ch.	Rel. +apart+ch.	Rel. +together +ch.	Married+no ch.	Married+ch.
Single+no ch.	0									
Single+ch.	2.000	0								
Changing rel. +no ch.	1.984	2.000	0							
Changing rel. +ch.	2.000	2.000	2.000	0						
Rel.+apart+no ch.	2	2	2	2	0					
Rel. +together+no ch.	2	2	2	2	2	0				
Rel.+apart+ch.	2	2	2	2	2	2	0			
Rel. +together+ch.	1.995	1.922	1.998	1.951	1.990	1.997	1.971	0		
Married+no ch.	1.985	2.000	1.974	2.000	2	2	2	2.000	0	
Married+ch.	1.984	1.998	1.976	2	1.975	1.958	1.996	1.984	1.986	0

Distance matrix

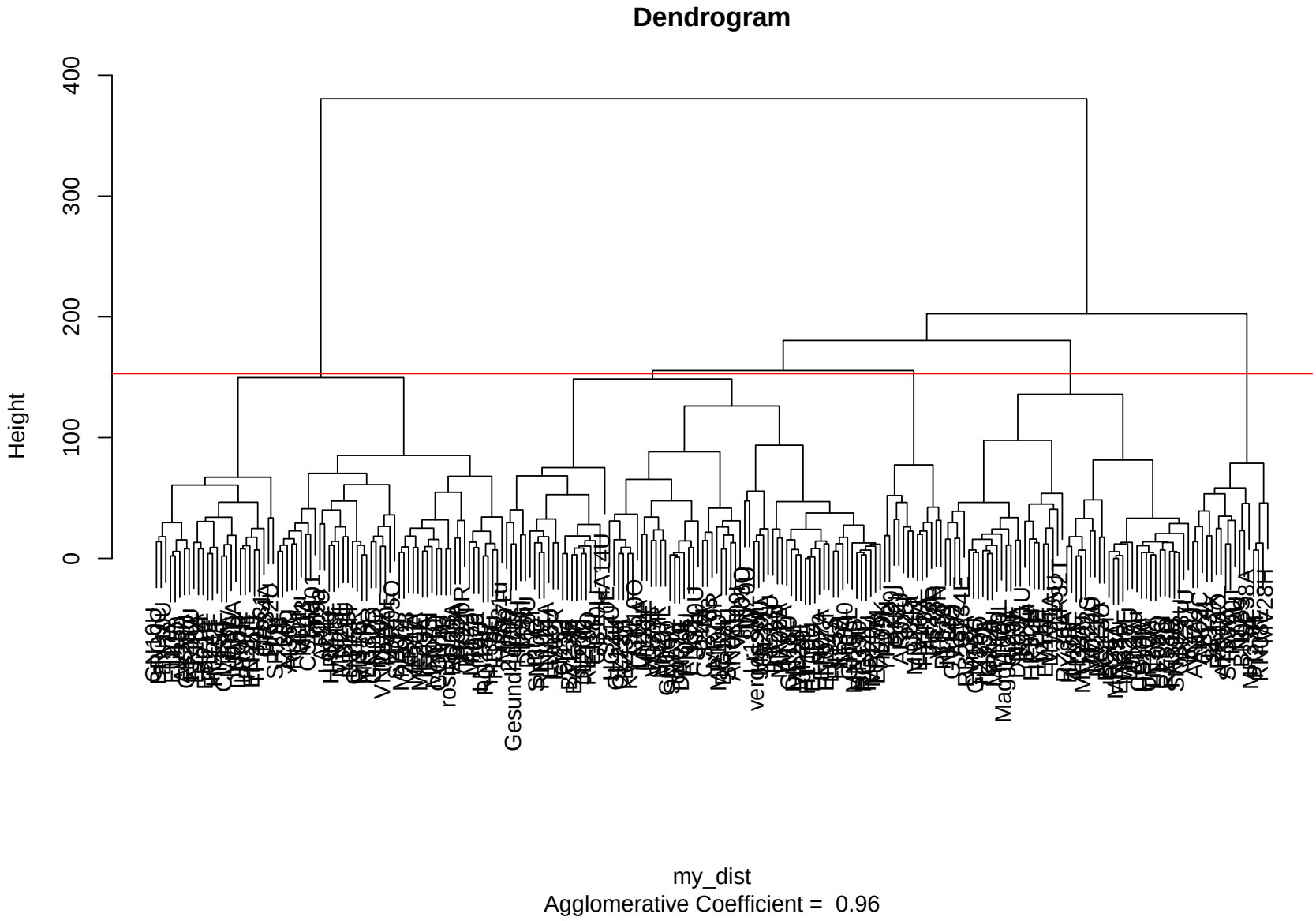
- Given $x, y \in X$ two sequences of interest. There different mappings from $T : X \to X$ such that $T(x) = y$.
- T is composed of elements (operations) that can be insertion, deletion or substitution.
- There is a cost associated with each operation: The substitution cost are given by the cost matrix and insertion/deletion costs are set in a way that reduces/increases the importance of time shifts (low/high).
- The distance between x and y is given by the lower cost mapping.

Clustering (i)

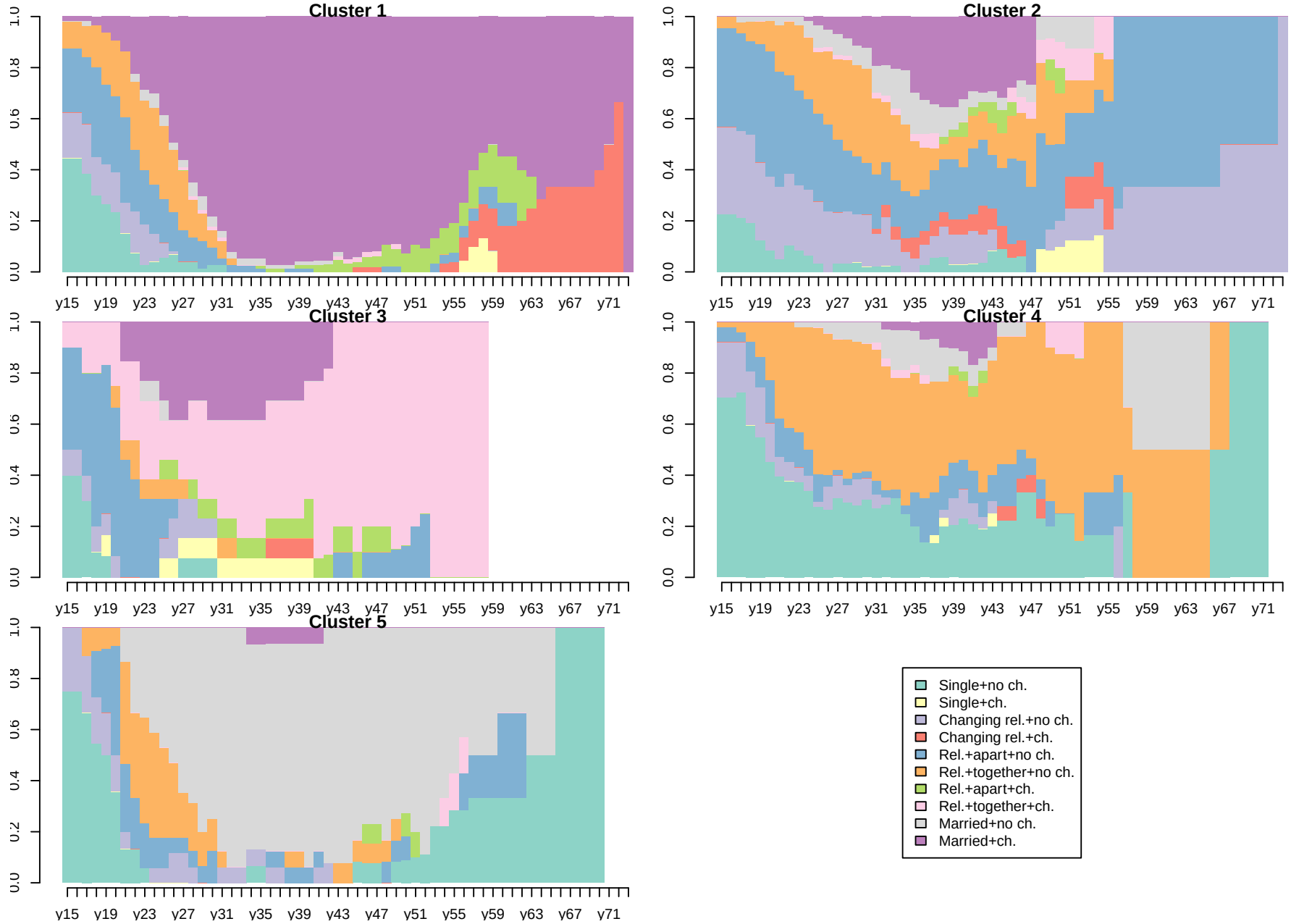
- Hierarchical method: Agglomerative Nesting (AGNES).
- At the beginning each individual is a cluster and, at every step, the closest clusters are merged together.
- Distance between two clusters is the average of the distances between the points in one cluster and the points in the other cluster.

Clustering (ii)

Dendrogram:



Clustering (iii)



Descriptive statistics of personality scores

Personality trait	Min	Max	Average	Std. deviation
Agreeablenes	1.50	5.0	3.43	0.76
Conscientiousness	1.75	5.0	4.15	0.57
Extraversion	1.50	5.0	3.63	0.82
Neuroticism	1.00	4.5	2.62	0.77
Openness	1.80	5.0	3.89	0.70

Average personality scores by cluster

Cluster	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Cluster 1	3.82	3.58	4.29	2.50	3.90
Cluster 2	3.51	3.37	4.10	2.77	4.00
Cluster 3	3.38	3.54	4.31	2.35	3.83
Cluster 4	3.58	3.38	4.12	2.61	3.79
Cluster 5	3.64	3.17	4.09	2.77	4.44

Subjective description of the clusters:

- Cluster 1: Married with children then divorced/widowed
- Cluster 2: Sequences with more changes (unstable)
- Cluster 3: Younger, not married with children
- Cluster 4: Not married w/o children
- Cluster 5: Married w/o children

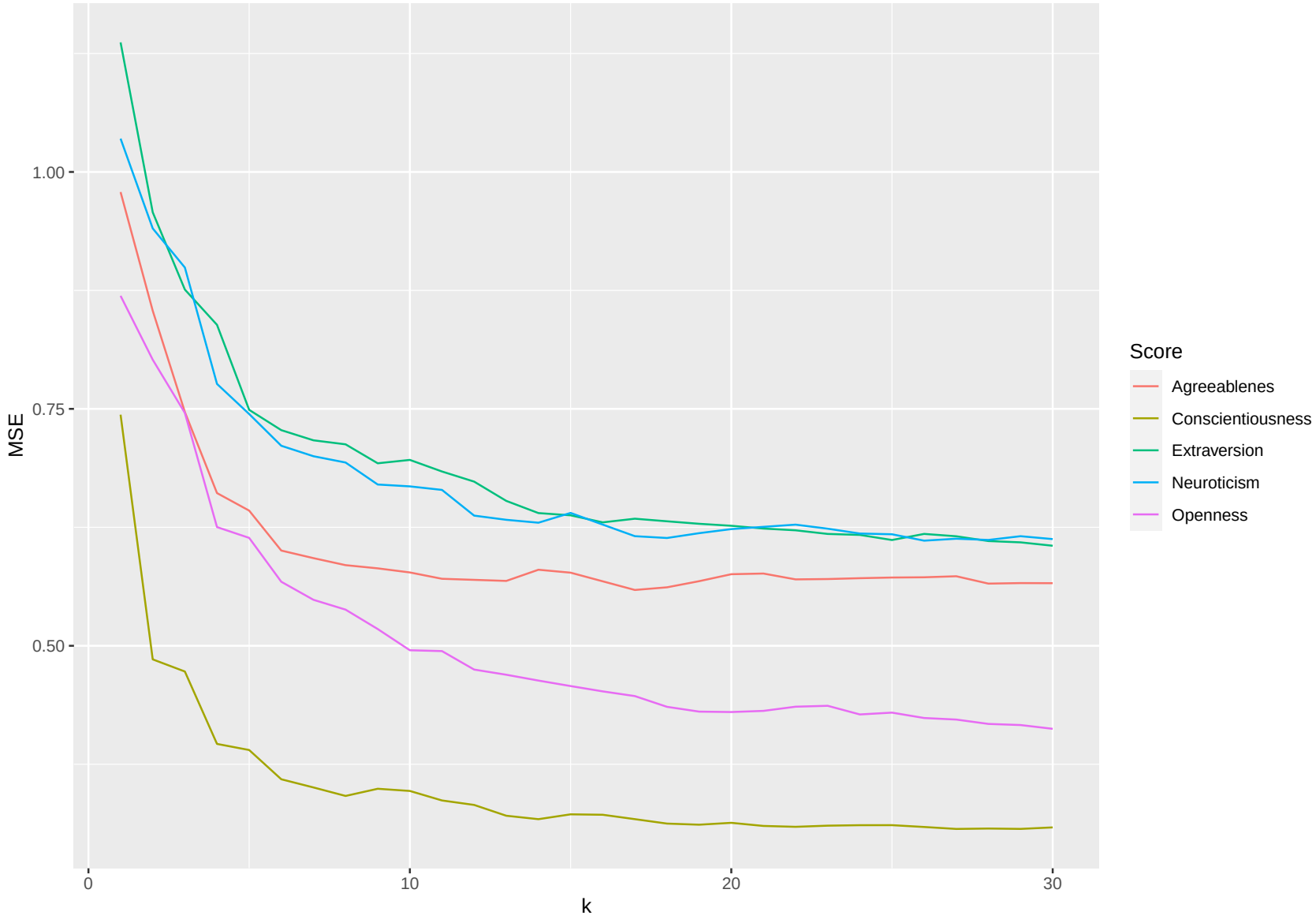
k-Nearest Neighbors (kNN) algorithm

- It's a non-parametric method.
- Choose the k nearest samples to an individual (distance matrix).
- Calculate the average of the variable of interest with the k samples $\rightarrow$ prediction.
- Use a measure such as MSE to select the optimal k.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$

where n is the number of data points considered, Y_i is the observed value and $\hat{Y}_i$ is the predicted value.

k-Nearest Neighbors



What's next?

- Use the kNN predictions to tune the parameters used in the specification of the cost matrix (e.g. indel cost, transition cost calculation)
- Try other prediction methods (e.g. distance-based linear models)

References

- Sequence Analysis: New Methods for Old Ideas - A. Abbott (1995)
- Optimal Matching Analysis: A Methodological Note on Studying Career Mobility - T. W. Chan (1995)
- Analyzing Sequence Data: Optimal Matching in Management Research - T. Biemann & D. K. Datta (2013)
- Analyzing and Visualizing State Sequences in R with TraMineR - A. Gabadinho, G. Ritschard, N. S. Müller, M. Studer (2011)