Analysis of the relationships history of women 40+

Adriana Clavijo Daza   Serena Lozza (Fiacco), PhD

Statistics and Data Science Master's, Universität Bern

ARTORG Center for Biomedical Engineering Research, Universität Bern

2022-12-12

Motivation

Understand the differences in the romantic relationships history of a group of women.

Women 40+ Healthy Aging Study (i)

Universität Zürich UZH

Dynamics of Healthy Aging

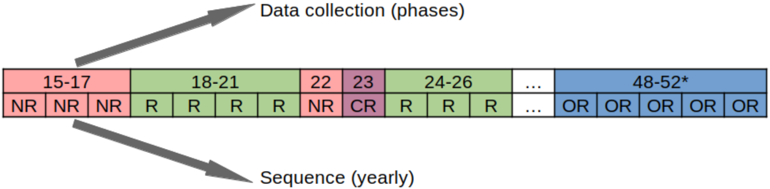
- Data from 250 individuals collected between June 2017 and February 2018.
- Psychometric instrument to obtain information about the history of romantic relationships of women aged between 40 and 75 years.

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.
- The data of the phases is then used to build a yearly sequence.

Data example

Consider the relationship status:



Data collection (phases)

15-17			18-21				22	23	24-26			...	48-52*				
NR	NR	NR	R	R	R	R	NR	CR	R	R	R	...	OR	OR	OR	OR	OR

Sequence (yearly)

*Current age

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

Research question

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

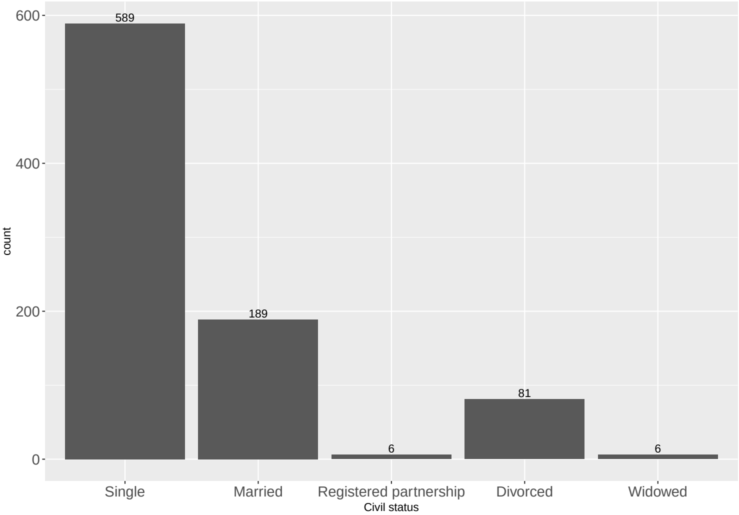
Progress

- Data cleaning/pre-processing
- Data exploration to find the variables to use (get a single sequence from several variables)
- Calculation and evaluation of cost matrix
- Hierarchical clustering
- Personality scores prediction using k-nearest neighbors

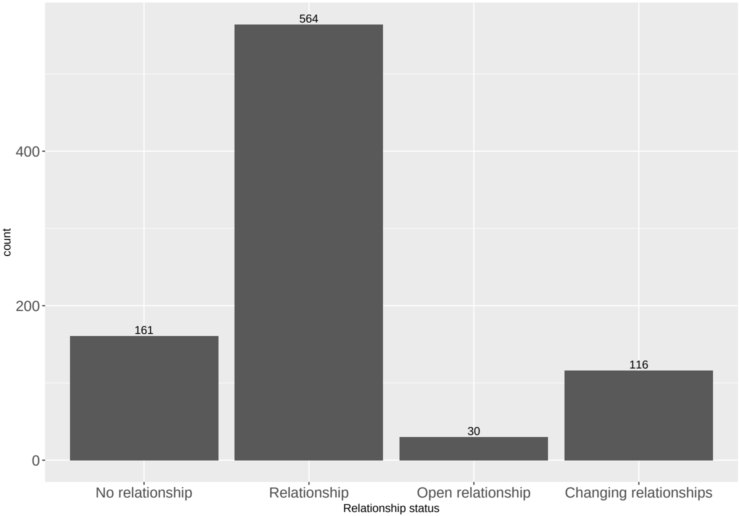
Data pre-processing

- 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.

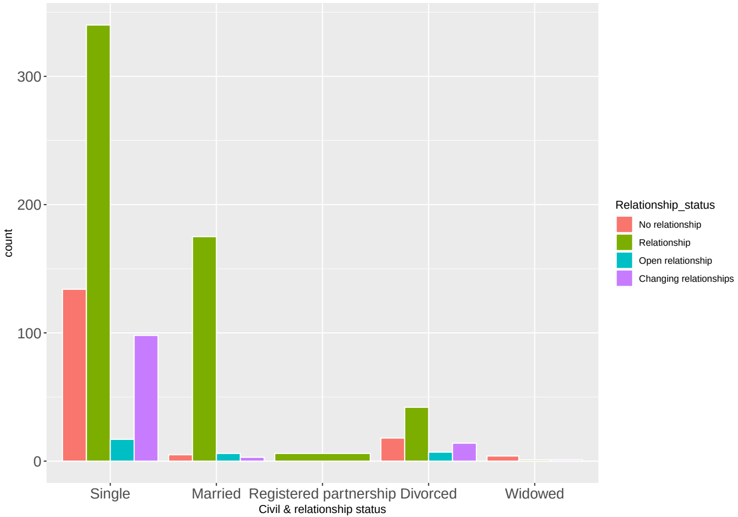
Frequency of phases - Civil status



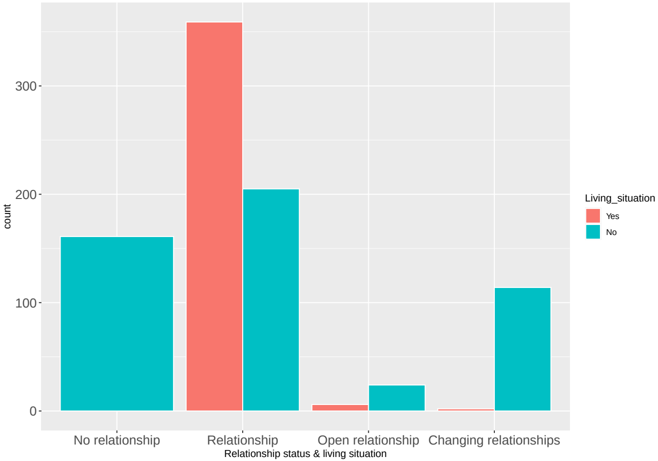
Frequency of phases - Relationship status



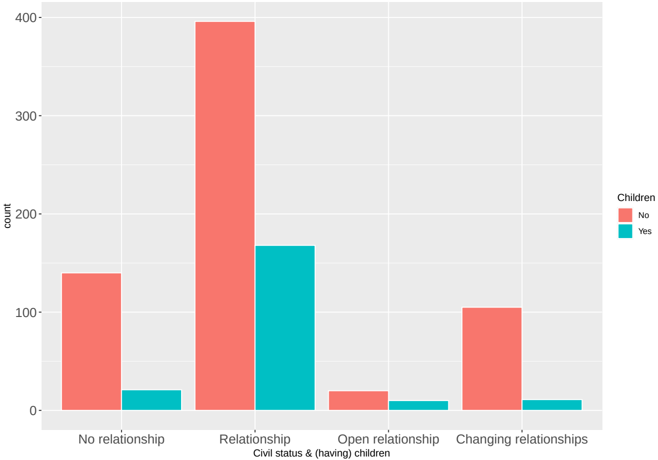
Frequency of phases - Civil & relationship status



Frequency of phases - Relationship status & living situation



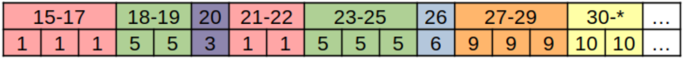
Frequency of phases - Civil status & (having) children



The instrument asked about the number of children in different phases but it will only be considered the presence/absence of children.

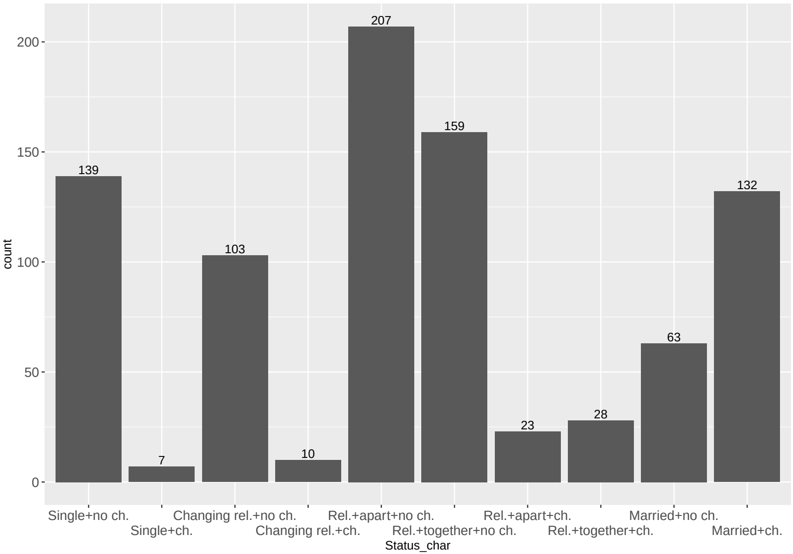
Considered states

- 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



15-17			18-19		20	21-22		23-25			26	27-29			30-*		...
1	1	1	5	5	3	1	1	5	5	5	6	9	9	9	10	10	...

Distribution of states



Optimal Matching Analysis (OMA)

- Technique used in social sciences for the comparison of sequences.
- Applications on life course and career path analysis.
- Given two sequences, it is possible to transform one sequence into another using a set of operations on the states: insert, delete and replace (*edit distance*).
- Numerical values are assigned to each of this operations and are defined in a **cost matrix**.
- As a result, pairwise distances between the sequences can be obtained to apply a clustering method.

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).

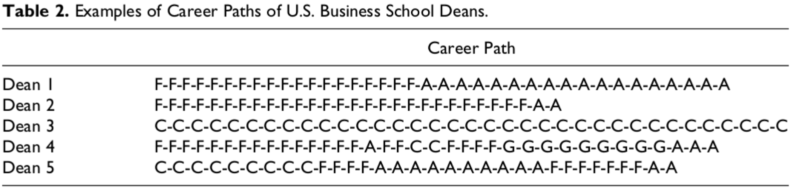


Table 2. Examples of Career Paths of U.S. Business School Deans.

	Career Path
Dean 1	F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-F-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-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
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:

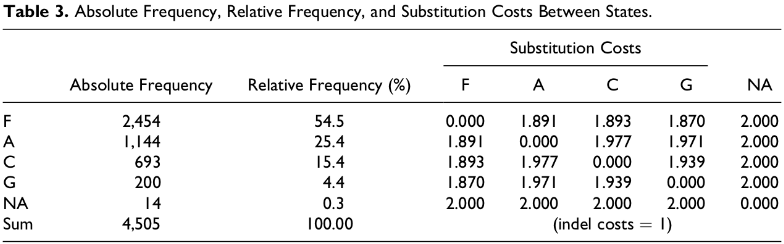


Table 3. Absolute Frequency, Relative Frequency, and Substitution Costs Between States.

			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:

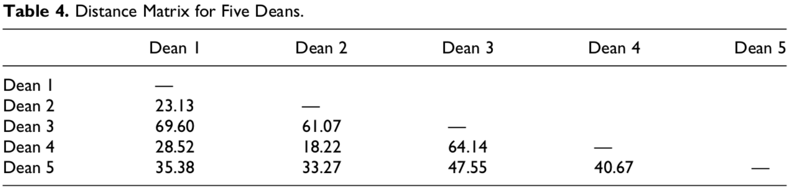
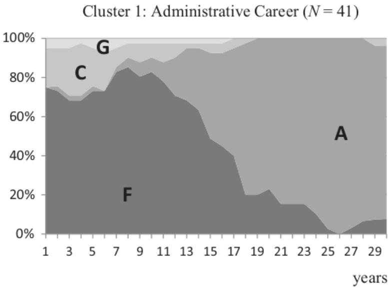


Table 4. Distance Matrix for Five Deans.

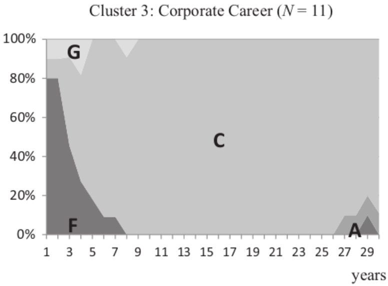
	Dean 1	Dean 2	Dean 3	Dean 4	Dean 5
Dean 1	—				
Dean 2	23.13	—			
Dean 3	69.60	61.07	—		
Dean 4	28.52	18.22	64.14	—	
Dean 5	35.38	33.27	47.55	40.67	—

Example (iv)

Two of the five resulting clusters:



Cluster 1: Administrative Career ($N = 41$)



Cluster 3: Corporate Career ($N = 11$)

Cost matrix (i)

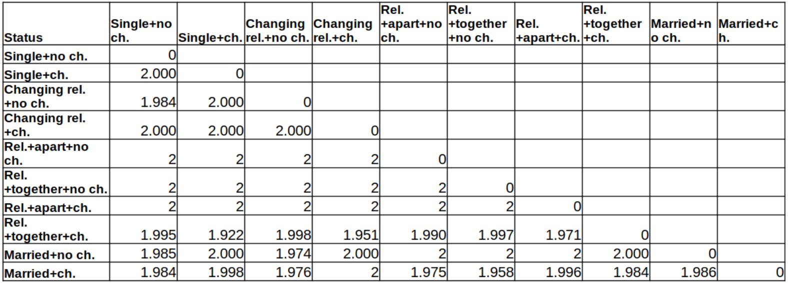
Using the R package `TraMineR` the cost matrix is calculated with transition rates between states.

Given a set of k states, say, $S = \{s_1, \ldots, s_k\}$, the substitution cost between states s_i and s_j, $1 \leq i, j \leq k$, is calculated as:

$$C(s_i, s_j) = c - P(s_i|s_j) - P(s_j|s_i)$$

where $P(s_i|s_j)$ is the probability of transition from state s_i in time t to s_j in time $t+1$ and c is a constant (set by default to $c = 2$ so that $0 \leq C(s_i, s_j) \leq 2)_,$.

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

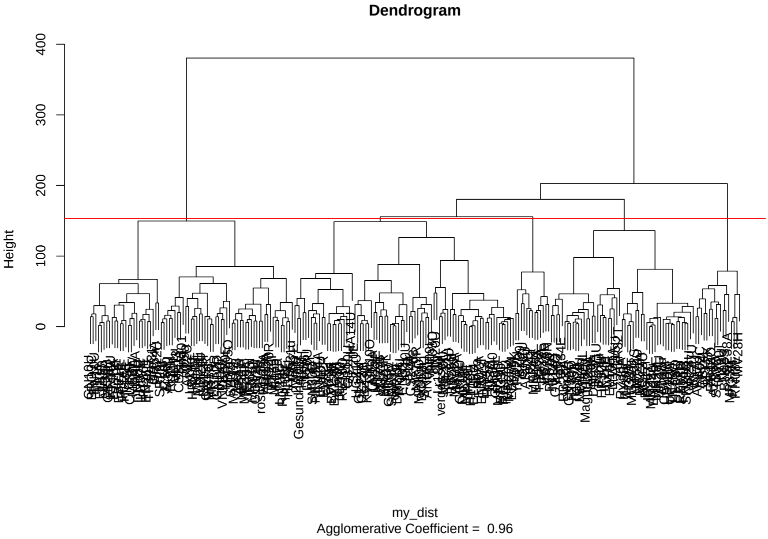
- Given $x, y \in X$ two sequences of interest. There different mappings from $T : X \to X$ such that $T(x) = y$.
- T is composed of elements (operations) that can be insertion, deletion or substitution.
- There is a cost associated with each operation: The substitution cost are given by the cost matrix 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)

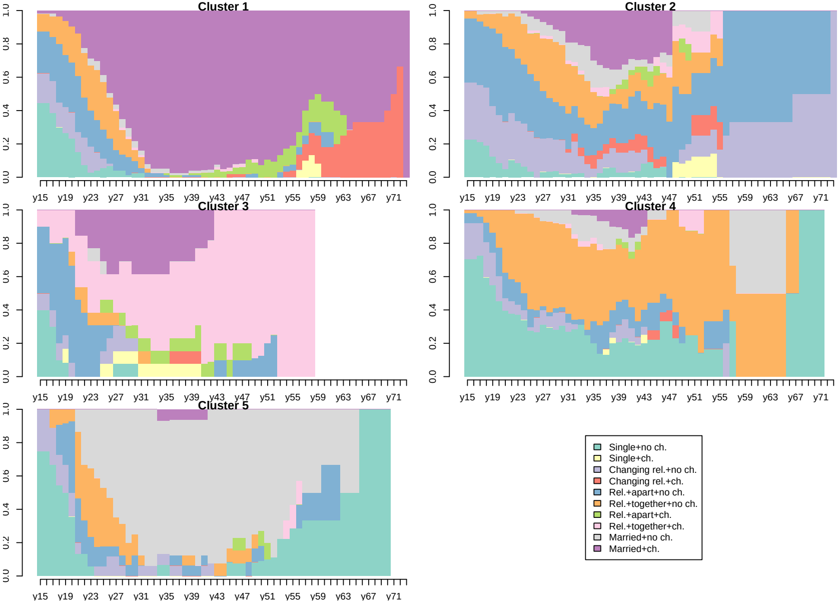
- Hierarchical method: Agglomerative Nesting (AGNES).
- At the beginning each individual is a cluster and, at every step, the closest clusters are merged together.
- Distance between two clusters is the average of the distances between the points in one cluster and the points in the other cluster.

Clustering (ii)

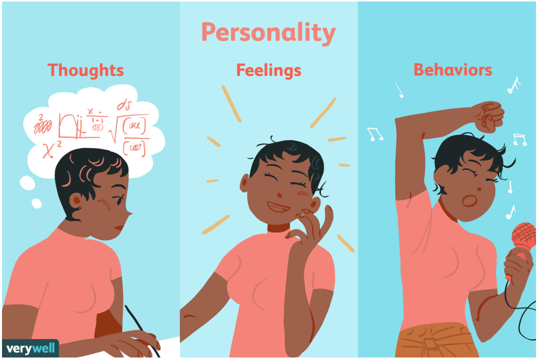
Dendrogram:



Clustering (iii)

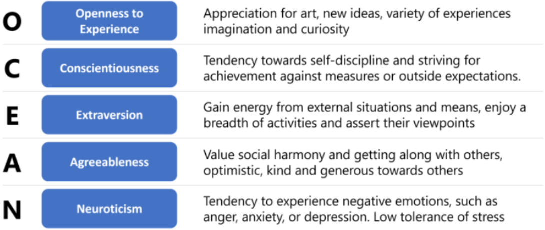


What is personality?



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“

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

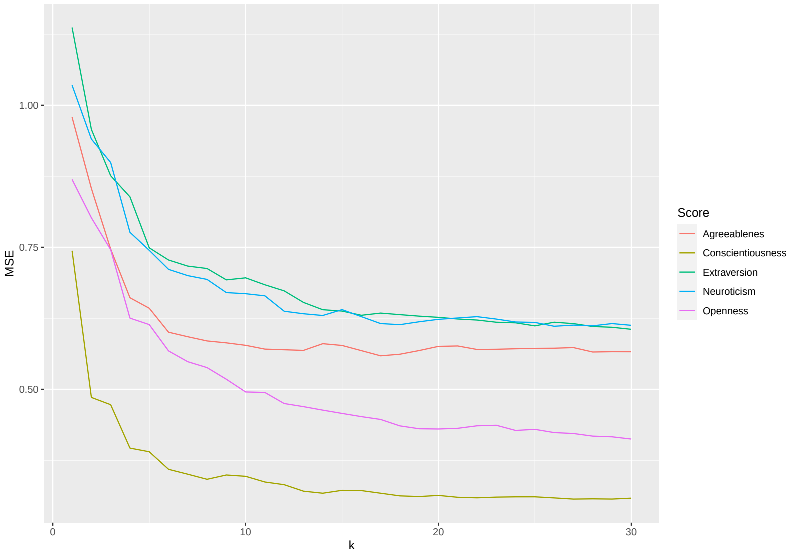
k-Nearest Neighbors (kNN) algorithm

- 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)