Categorical Sequence Analysis with Optimal Matching: An Application with Data from the 'Women 40+ Healthy Aging Study'

Adriana Clavijo Daza

Supervised by:
Prof. Dr. David Ginsbourger
Dr. Serena Lozza-Fiacco

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 explore the use of this information as a predictor for other psychosocial characteristics of interest.

Women 40+ Healthy Aging Study (i)

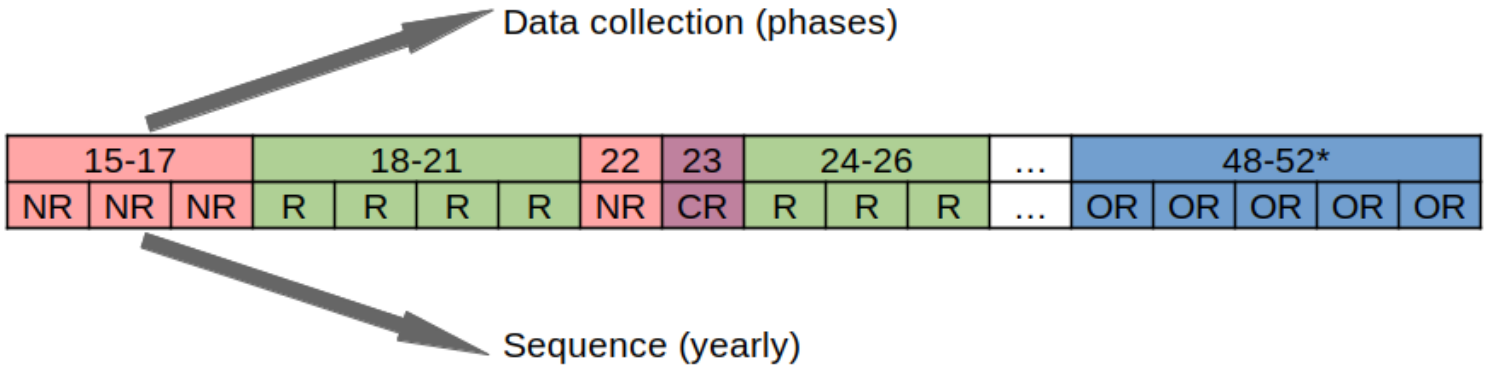
Universität Zürich[UZH]

Dynamics of Healthy Aging

- Data from 250 women (ages 40-75) collected between June 2017 and February 2018.
- Information about relationship phases starting from the age of 15 years until the current age obtained with a psychometric instrument.
- Phases defined by the start and end age
- For each phase: civil status, relationship status, living situation, children and quality of the relationship.
- Additional information collected, in particular, scores for personality traits.

Obtaining the sequences

Consider the relationship status: no relationship (NR), in a relationship (R), open relationship (OR), changing relationships (CR).

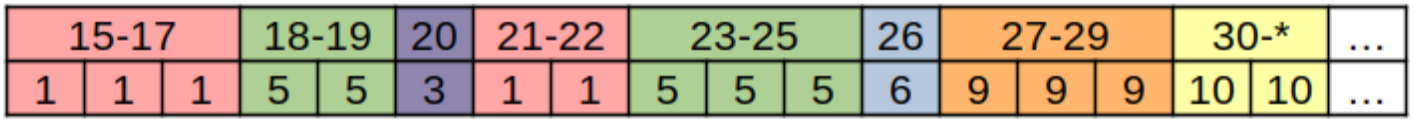


*Current age

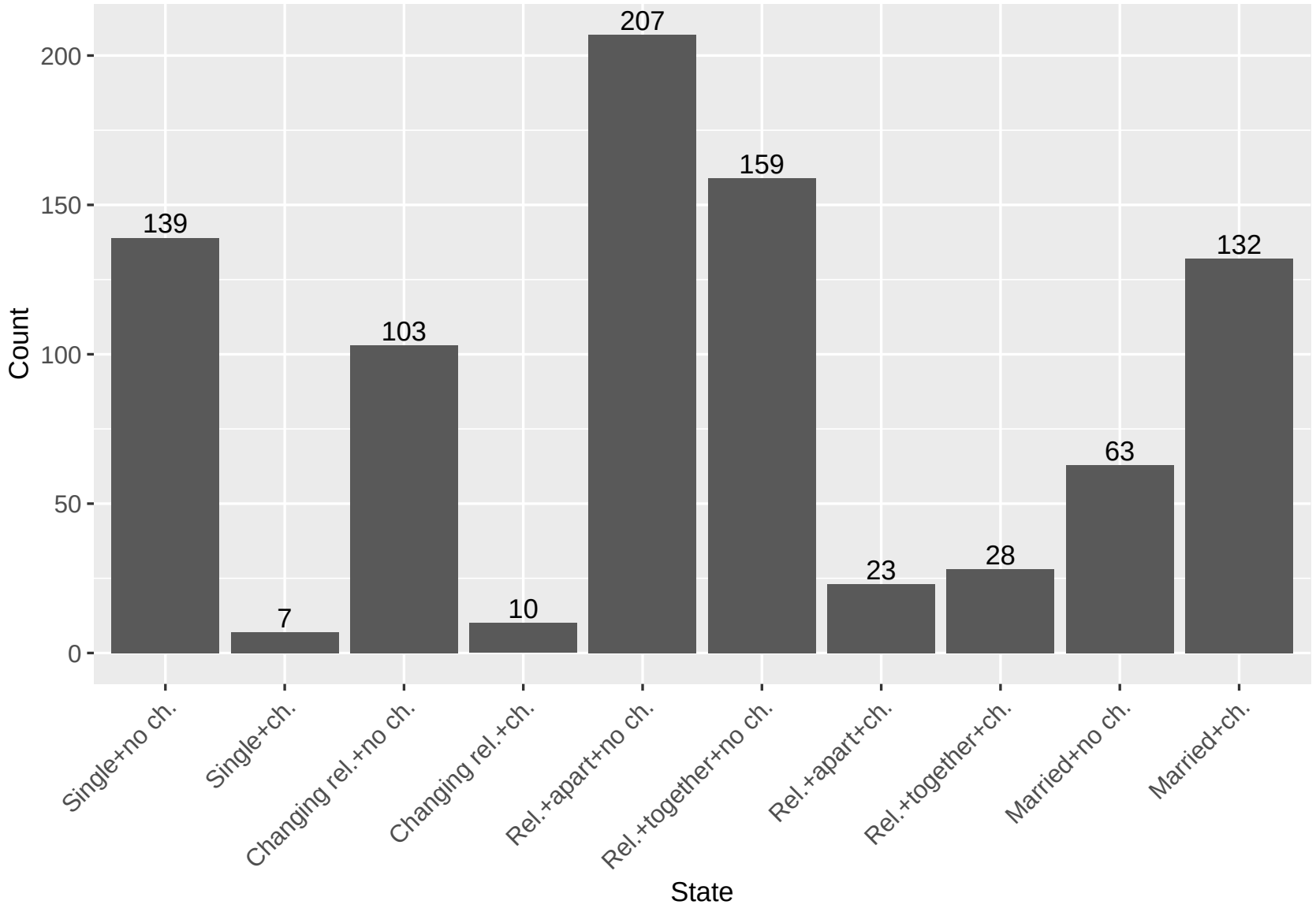
- Manual and automatic checks to identify inconsistent and incomplete records. Corrections based on secondary data source.
- Identification and selection of the variables that provide a wider perspective of the relationship situation at a given time.
- Sequence data available for 239 individuals.

Considered states

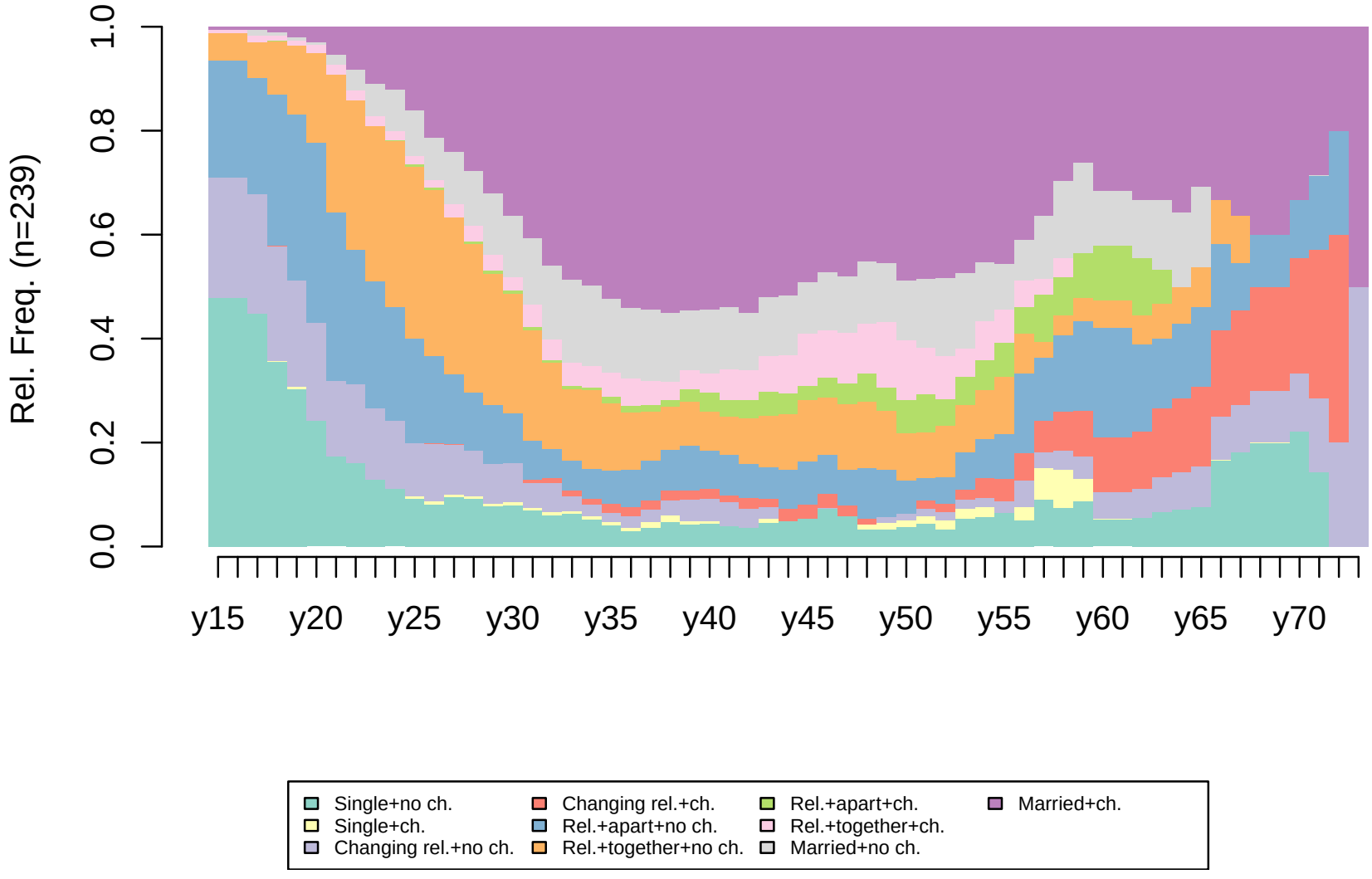
- 1 = Single + no children
- 2 = Single + children
- 3 = Changing relationships + no children
- 4 = Changing rel. + children
- 5 = Relationship + living apart + no children
- 6 = Relationship + living together + no children
- 7 = Relationship + living apart + children
- 8 = Relationship + living together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



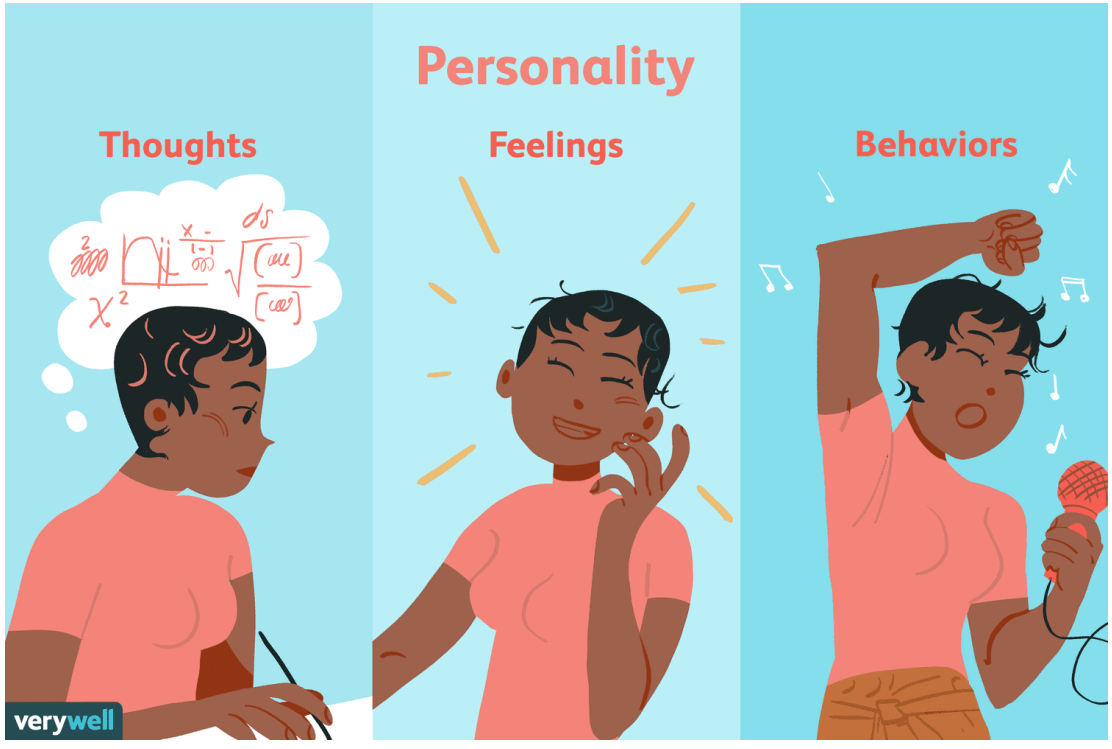
Overall distribution of states



Distribution of states by year



What is personality?

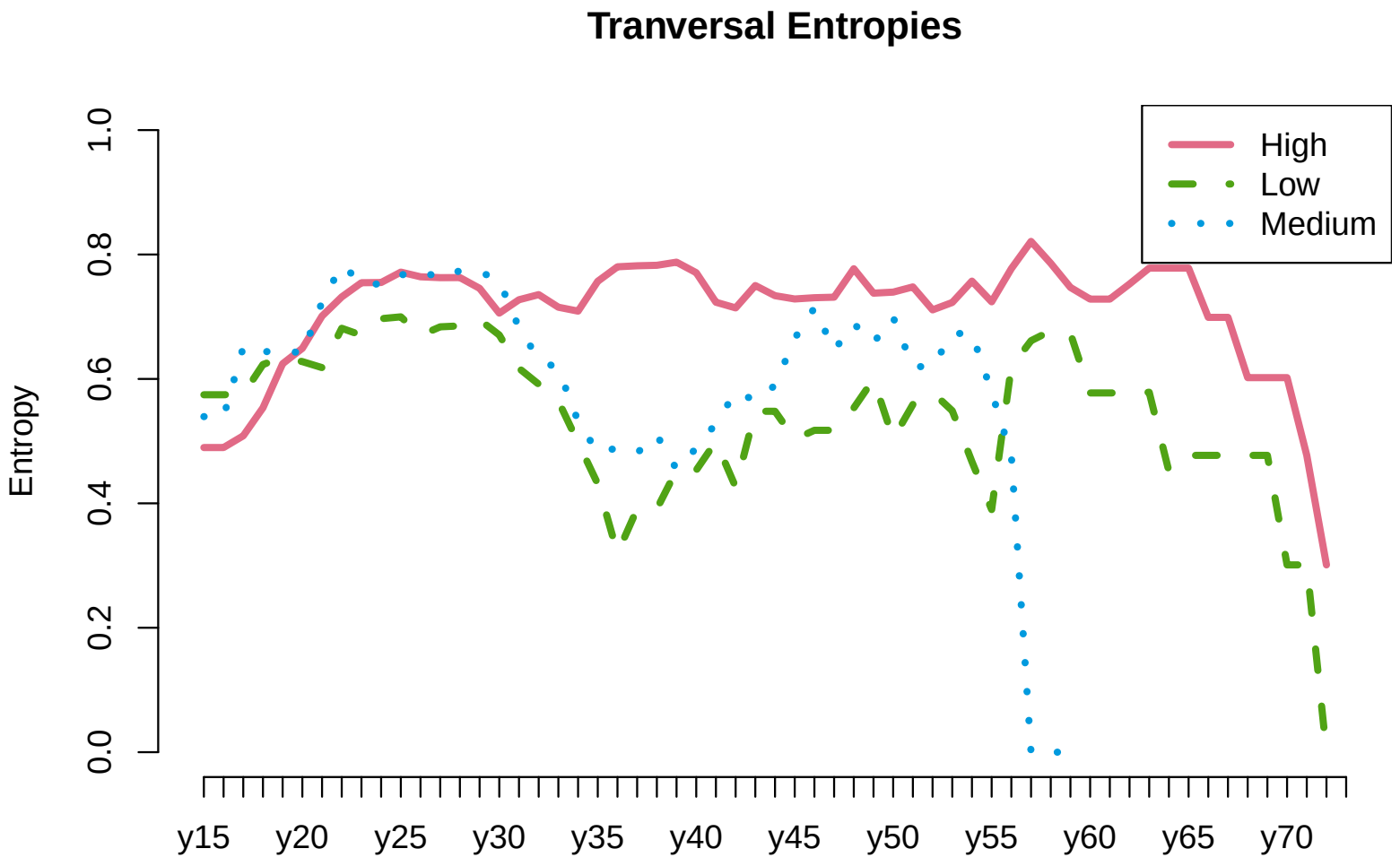


Personality refers to the enduring characteristics and behavior that comprise a person's unique adjustment to life, including major traits, interests, drives, values, self-concept, abilities, and emotional patterns.

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?

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.

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):

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

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

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)

Setup:

- $S =$ 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)

Initialization of F:

		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				

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

Example (iv)

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

S E N D with
A - N D or —> Optimal (equivalent) alignments
- A N D

- 2 matches: 0
- 1 mismatch: 3
- 1 gap: 2
- Total: 5

Normalization

Given $X, Y \in \mathbf{S}$ of length t_1 and t_2, respectively. Let $d(X, Y)$ be the distance between the sequences X and Y, t_{max} the length of the longest sequence in $\mathbf{S}$ and d_{max} the maximum distance between any pair of sequences in $\mathbf{S}$.

`TraMineR` provides the following options for normalization:

- `maxlength`: $d(X, Y)/t_{max}$
- `gmean`: $1 - \frac{d_{max} - d(X,Y)}{\sqrt{t_1 * t_2}}$
- `maxdist`: $d(X, Y)/d_{max}$

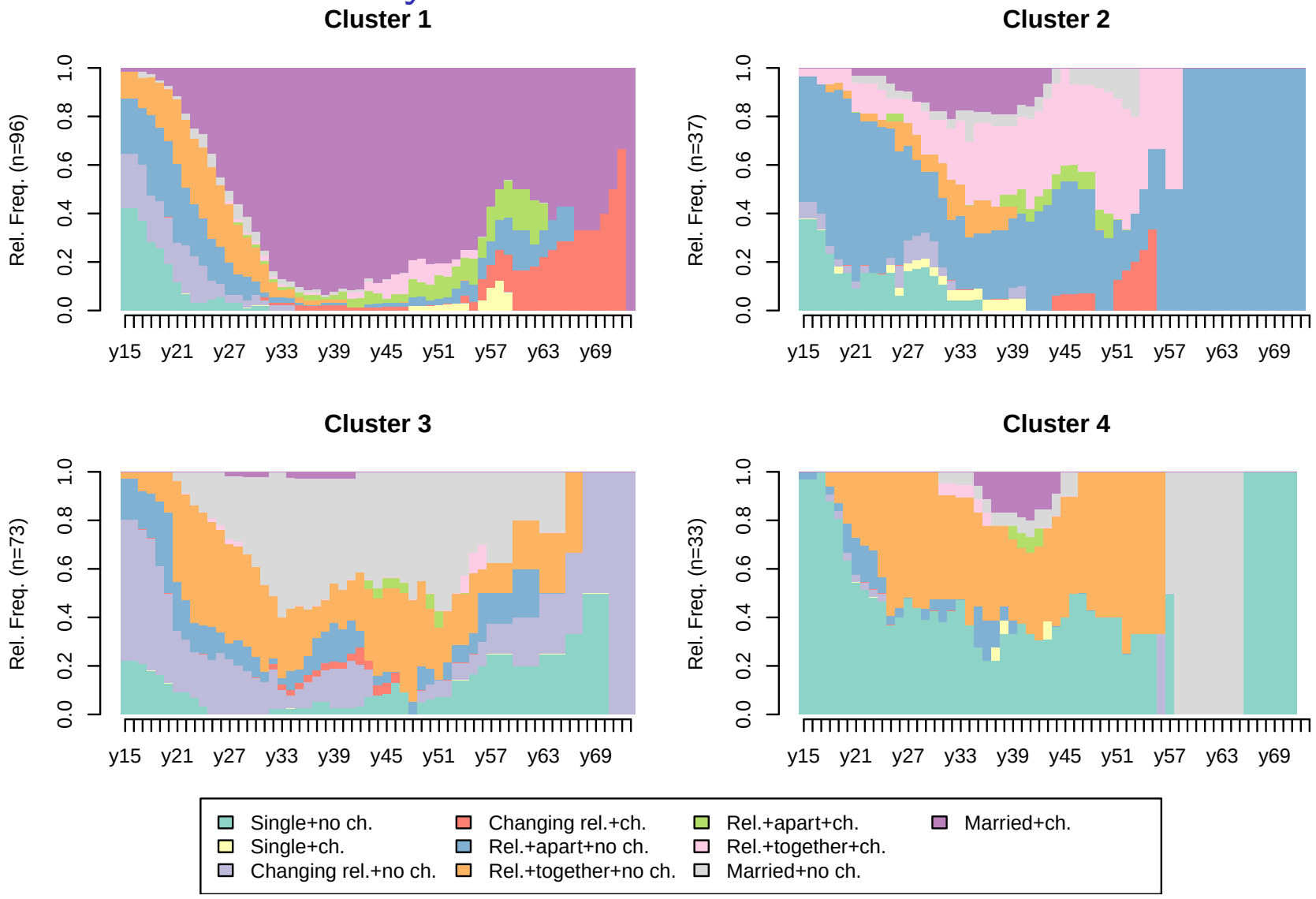
Cost matrix

Base setup: `TRATE` and `maxlenght`

State	1	2	3	4	5	6	7	8	9	10	NA
1	0.00	2.00	1.98	2.00	1.92	1.95	2.00	1.99	1.98	1.98	2
2	2.00	0.00	2.00	2.00	2.00	2.00	1.96	1.92	2.00	2.00	2
3	1.98	2.00	0.00	2.00	1.94	1.92	2.00	2.00	1.97	1.98	2
4	2.00	2.00	2.00	0.00	1.99	2.00	1.95	1.95	2.00	2.00	2
5	1.92	2.00	1.94	1.99	0.00	1.95	1.98	1.99	1.98	1.97	2
6	1.95	2.00	1.92	2.00	1.95	0.00	2.00	2.00	1.98	1.96	2
7	2.00	1.96	2.00	1.95	1.98	2.00	0.00	1.97	2.00	2.00	2
8	1.99	1.92	2.00	1.95	1.99	2.00	1.97	0.00	2.00	1.98	2
9	1.98	2.00	1.97	2.00	1.98	1.98	2.00	2.00	0.00	1.99	2
10	1.98	2.00	1.98	2.00	1.97	1.96	2.00	1.98	1.99	0.00	2
NA	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	0

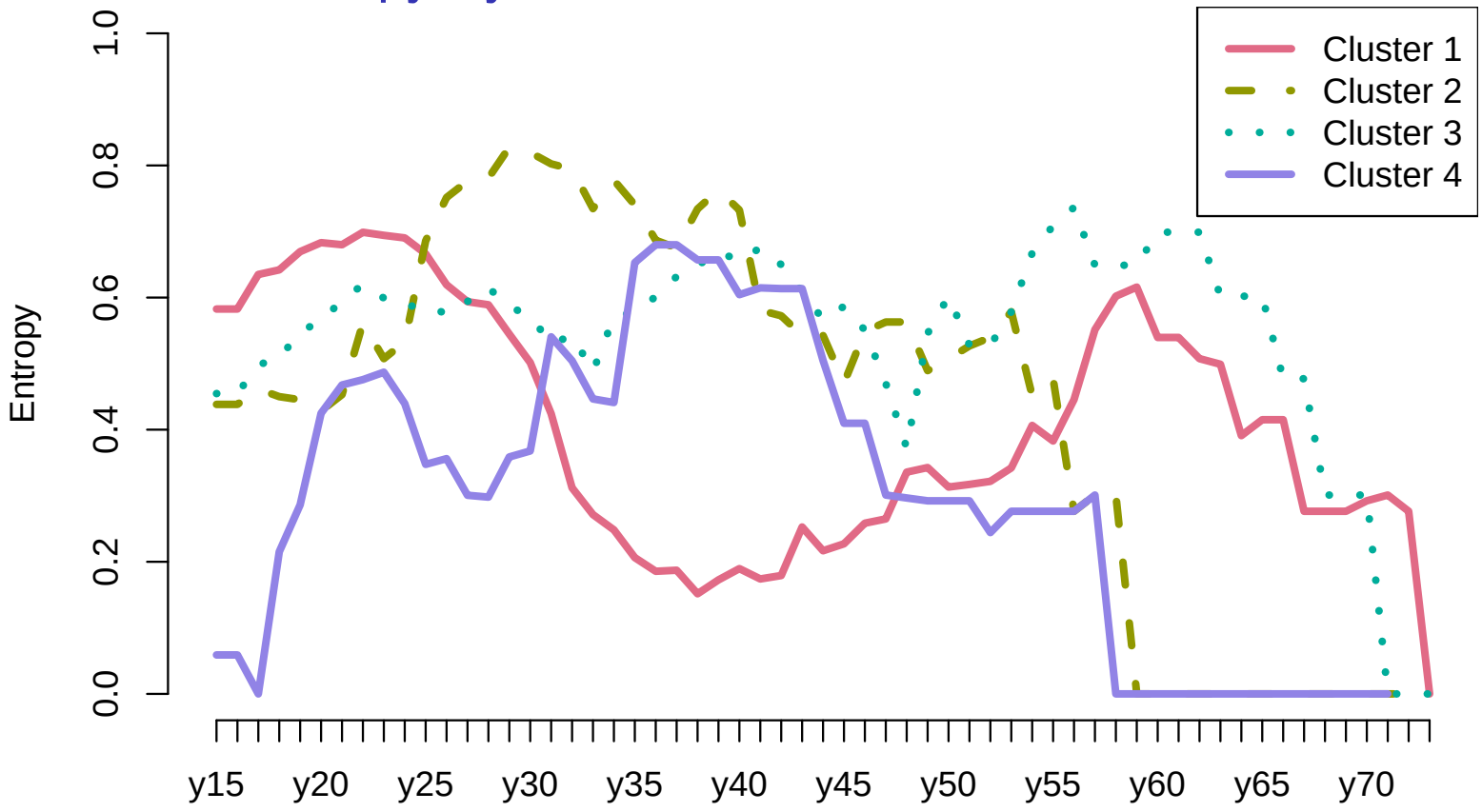
Clustering

State distribution by cluster



Application: clustering

Transversal entropy by cluster



Clustering

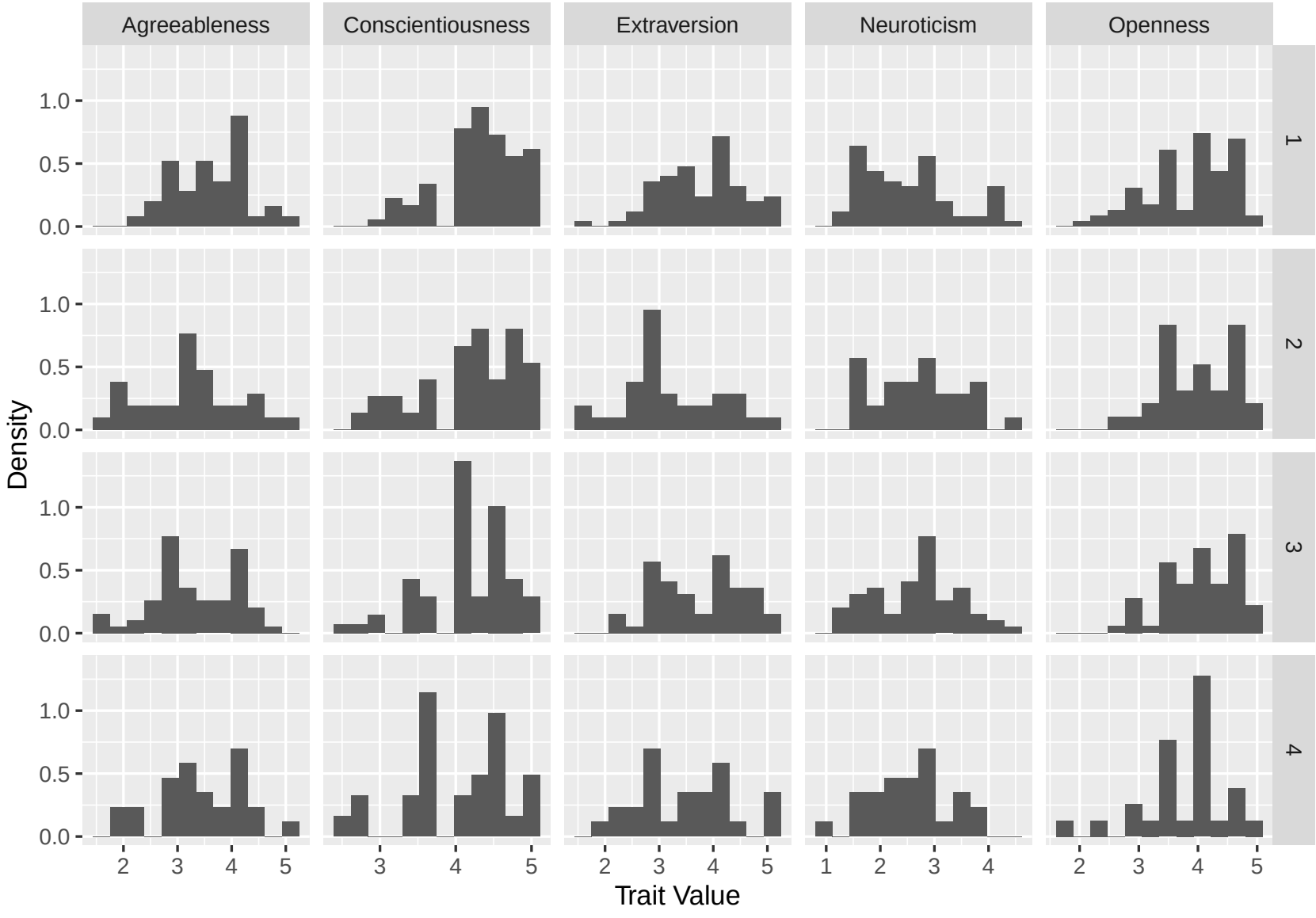
Characterization of the clusters

Cluster	n
1	96
2	37
3	73
4	33

- Cluster 1: Married young and had children
- Cluster 2: Often in relationships but not married
- Cluster 3: Older, mostly married or in long relationship without children
- Cluster 4: Younger, single or in a relationship without children

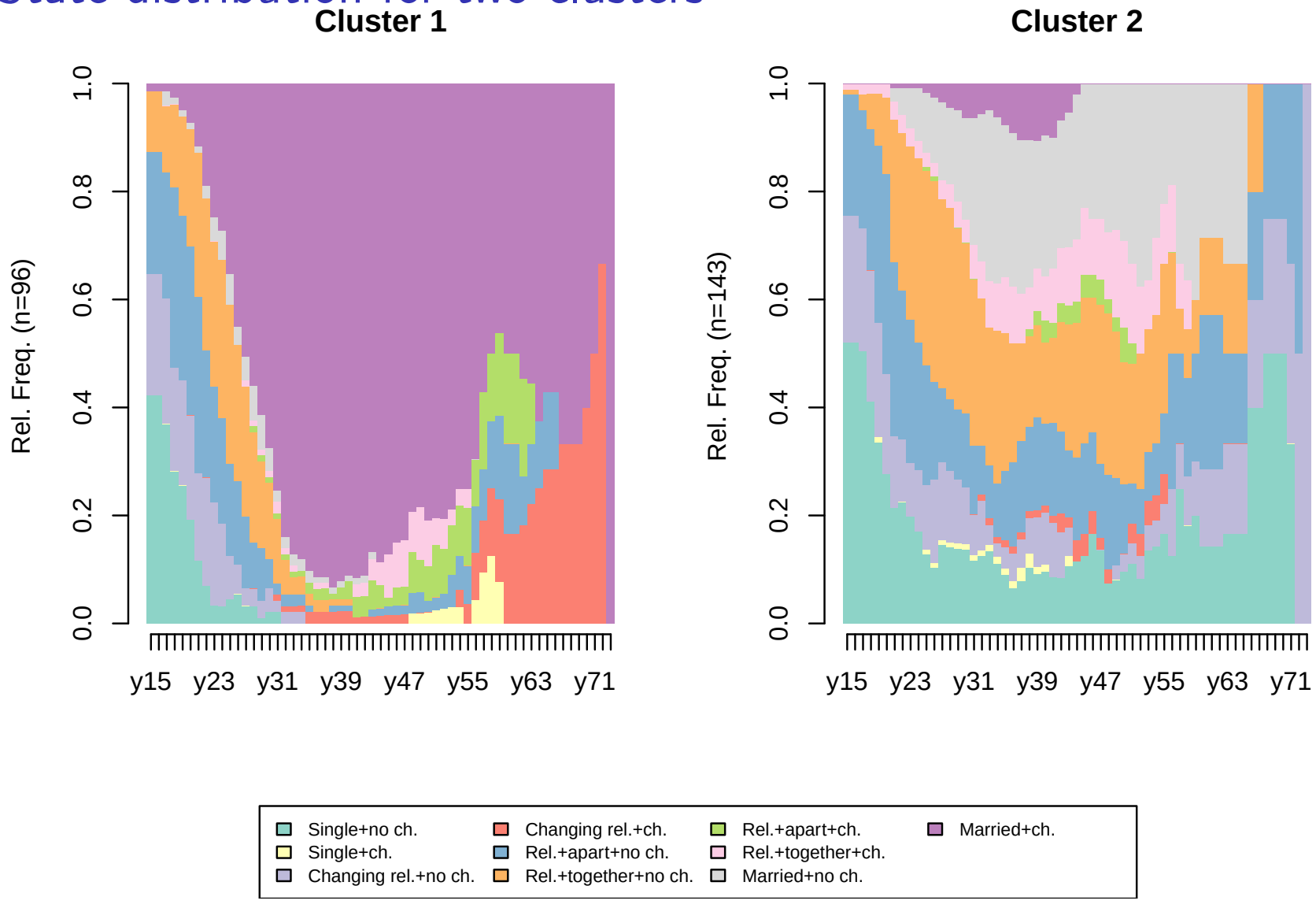
Clustering

Personality scores distribution by cluster



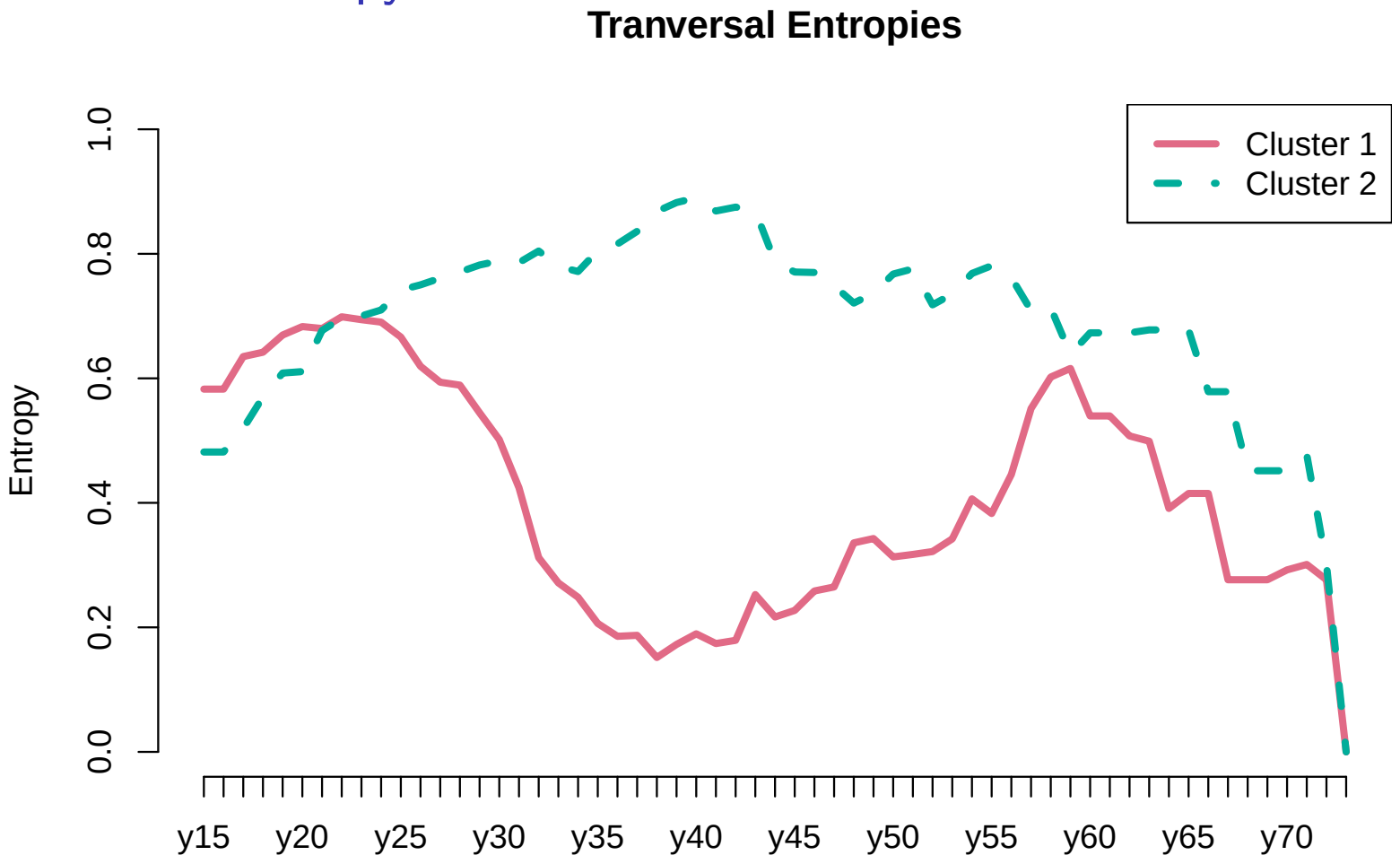
Clustering

State distribution for two clusters



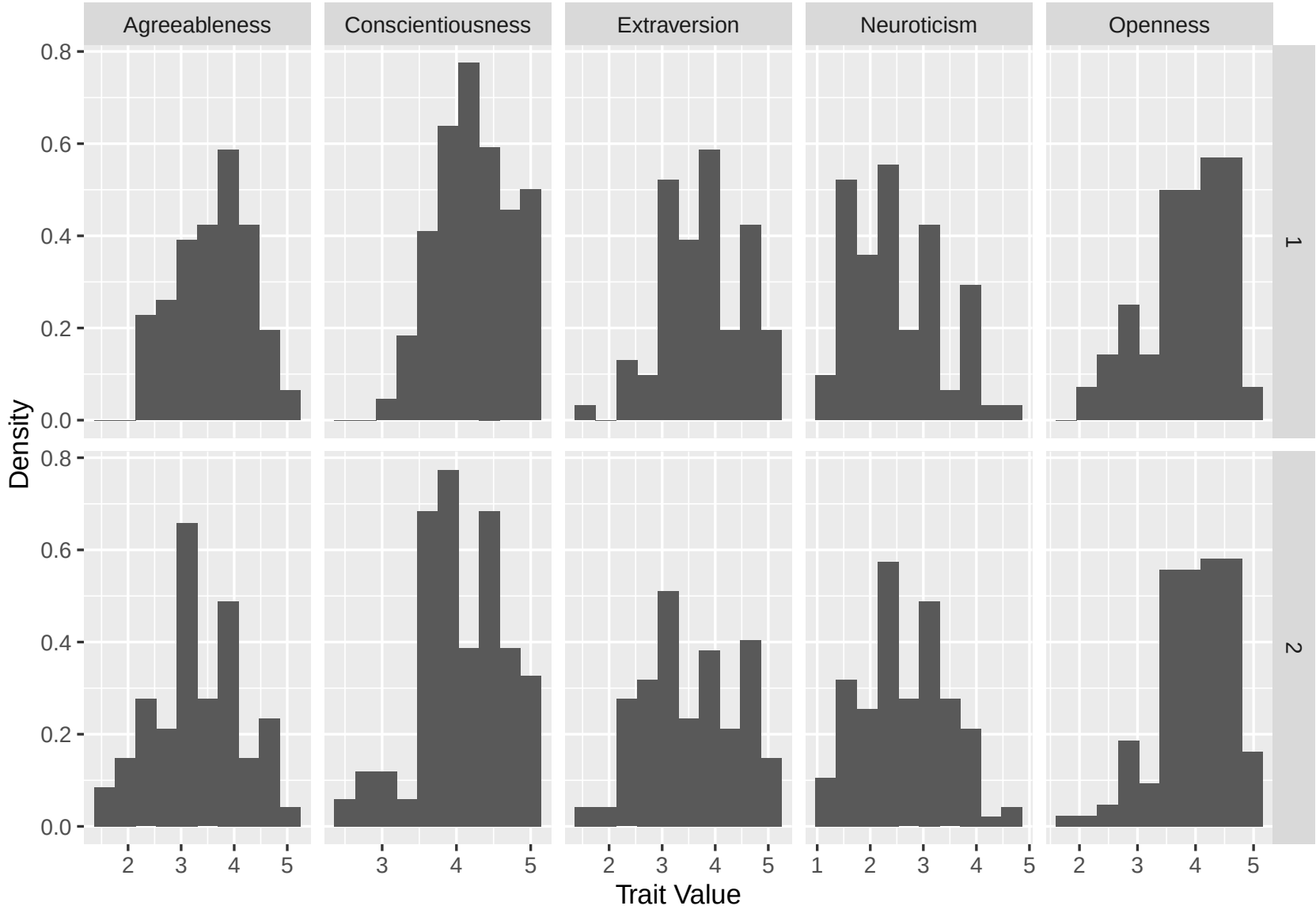
Application: clustering

Transversal entropy for two clusters



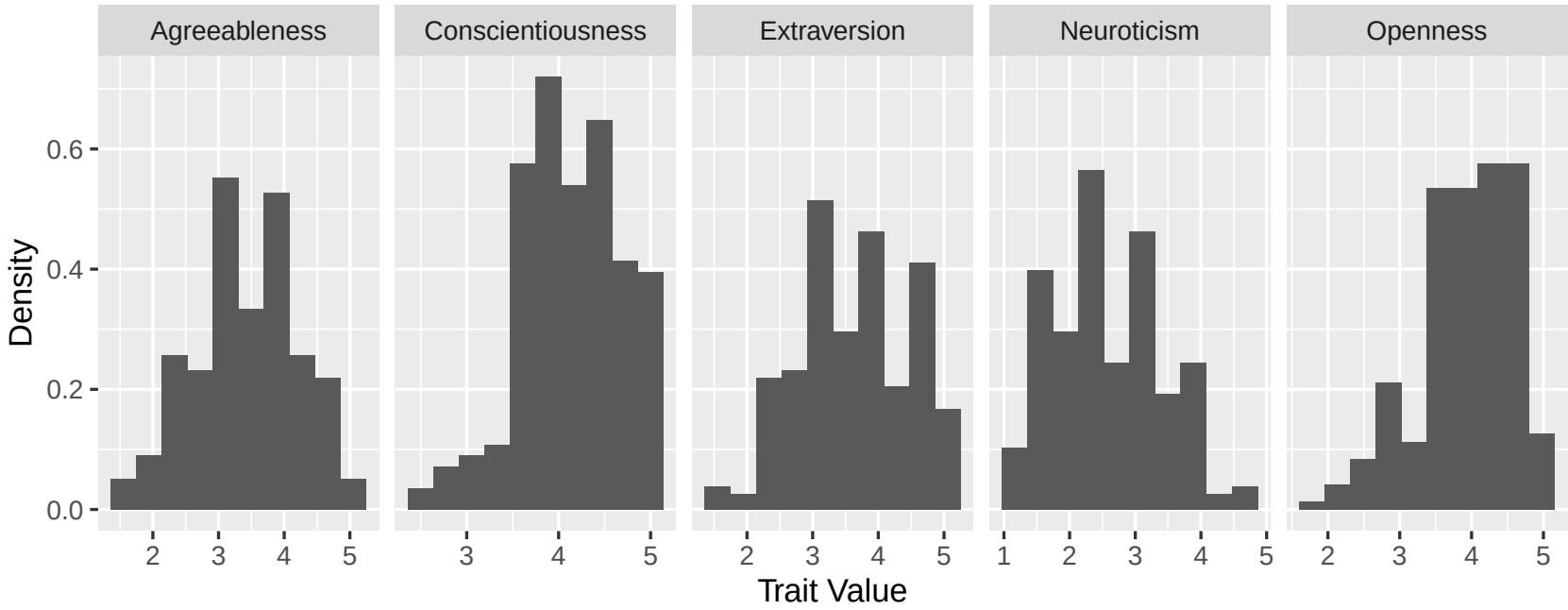
Application: clustering

Personality scores distribution for two clusters



Descriptive statistics for personality scores

Personality trait	Min	Max	Average	Std. deviation
Agreeableness	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



k-nearest neighbors (kNN)

- Non-parametric method for prediction.
- Training set $\mathcal{D} = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ of n labeled data points, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathcal{Y}$ (a finite set of for classification or a continuous range of values for regression).
- Predict the label or value (y_{n+1} unknown) for x_{n+1} by finding the k training data points closest to x_{n+1} and taking a majority vote of their labels (for classification) or averaging the values of Y (for regression).
- To compare the performance of different values of k, we calculate the mean squared error (MSE) in a test set of size m.

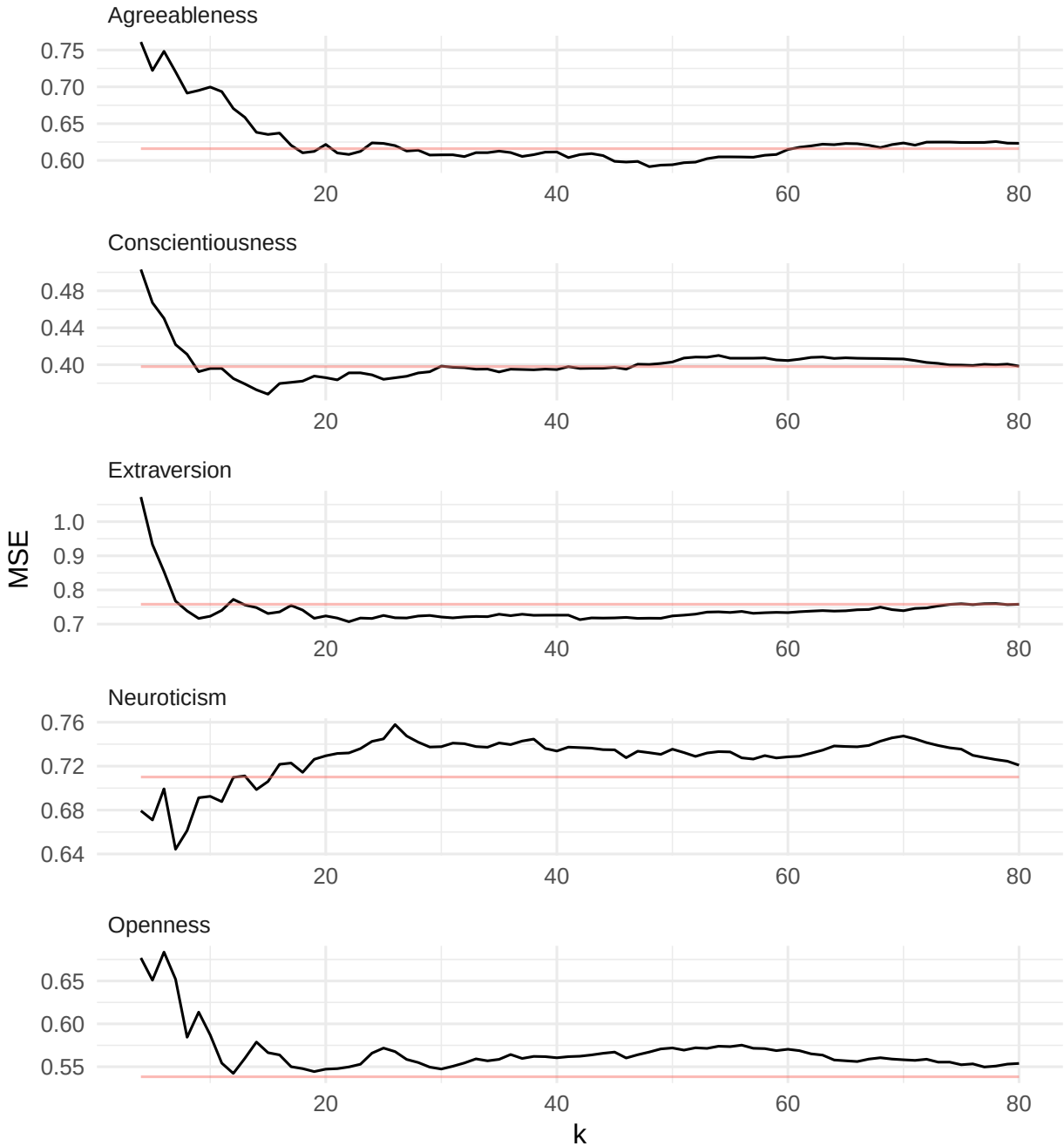
$$\text{MSE} = \frac{1}{m} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where y_i is the observed value and $\hat{y}_i$ is the predicted value via kNN.

Prediction of personality scores (i)

- Only individuals who have available personality scores
- New sample size: 200 individuals.
- Data split into two subsets: train (70%) and test (30%).
- Evaluate the MSE of the predictions in the test set only using the data from the nearest neighbors available in the train set.

Prediction of personality scores (ii)



Prediction of personality scores (iii)

- Method for calculation of cost matrix
- Constant c for cost matrix obtained from transition rates
- Normalization of distances
- Cost of transition from/to NA
- Limit range of the sequence

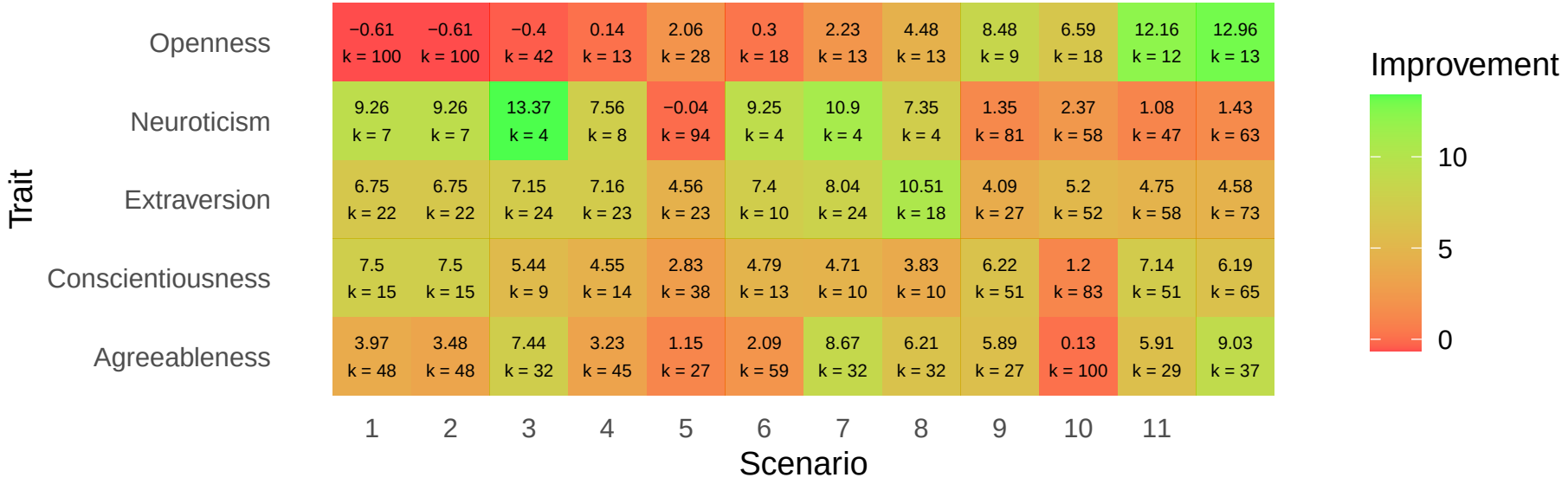
Prediction of personality scores

Scenarios considered

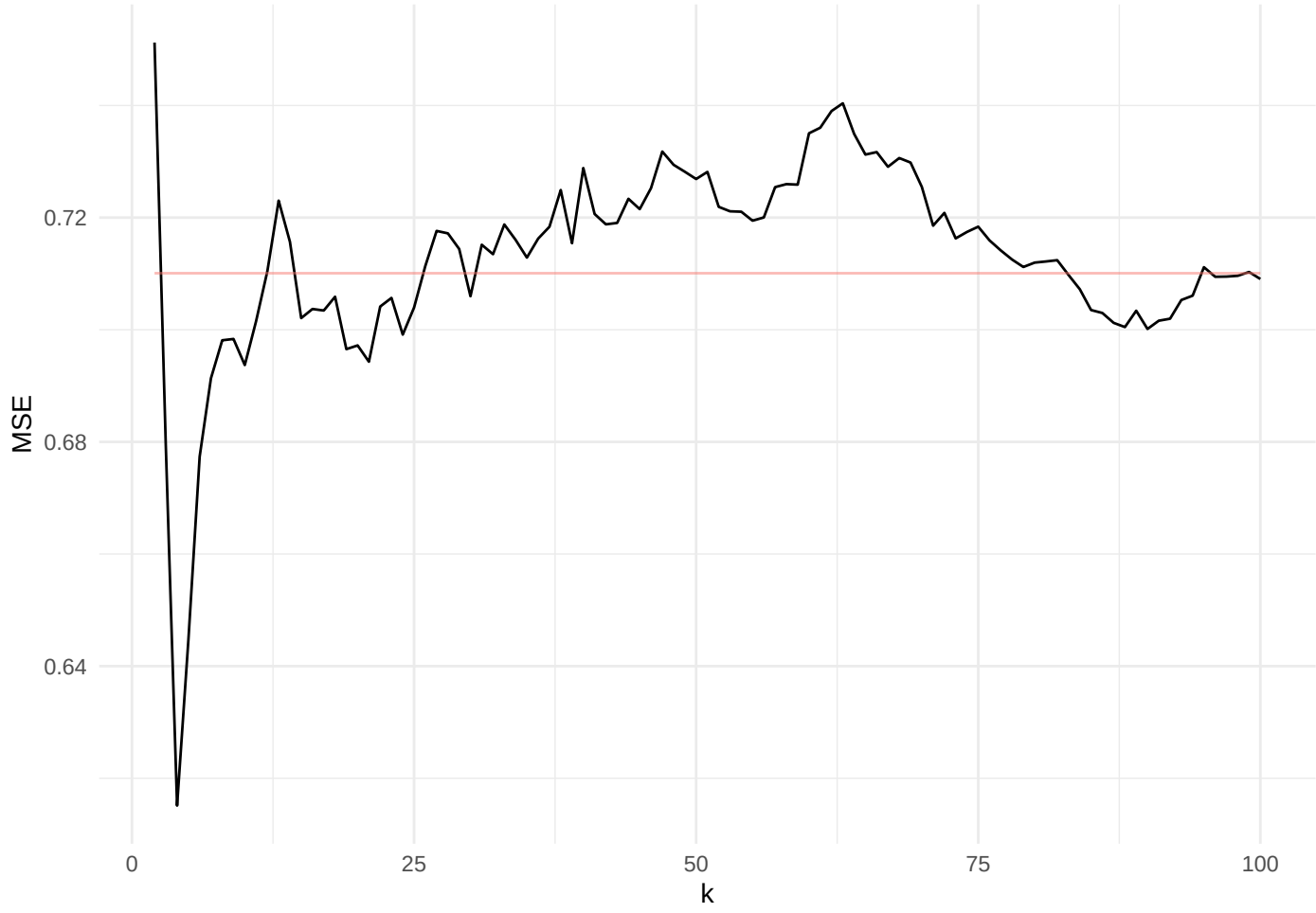
	cm_method	dm_norm	constant	missing_cost	min_age	max_age
1	TRATE	maxlength	NULL	NULL	NULL	NULL
2	TRATE	maxlength	1.998497	NULL	NULL	NULL
3	TRATE	gmean	NULL	NULL	NULL	NULL
4	FUTURE	maxlength	NULL	NULL	NULL	NULL
5	INDELS	maxlength	NULL	NULL	NULL	NULL
6	INDELSLOG	maxlength	NULL	NULL	NULL	NULL
7	FUTURE	gmean	NULL	NULL	NULL	NULL
8	FUTURE	gmean	NULL	1	NULL	NULL
9	FUTURE	gmean	NULL	NULL	20	55
10	FUTURE	gmean	NULL	NULL	20	40
11	FUTURE	gmean	NULL	1	20	55
12	FUTURE	gmean	NULL	0.5	20	55

Prediction of personality scores

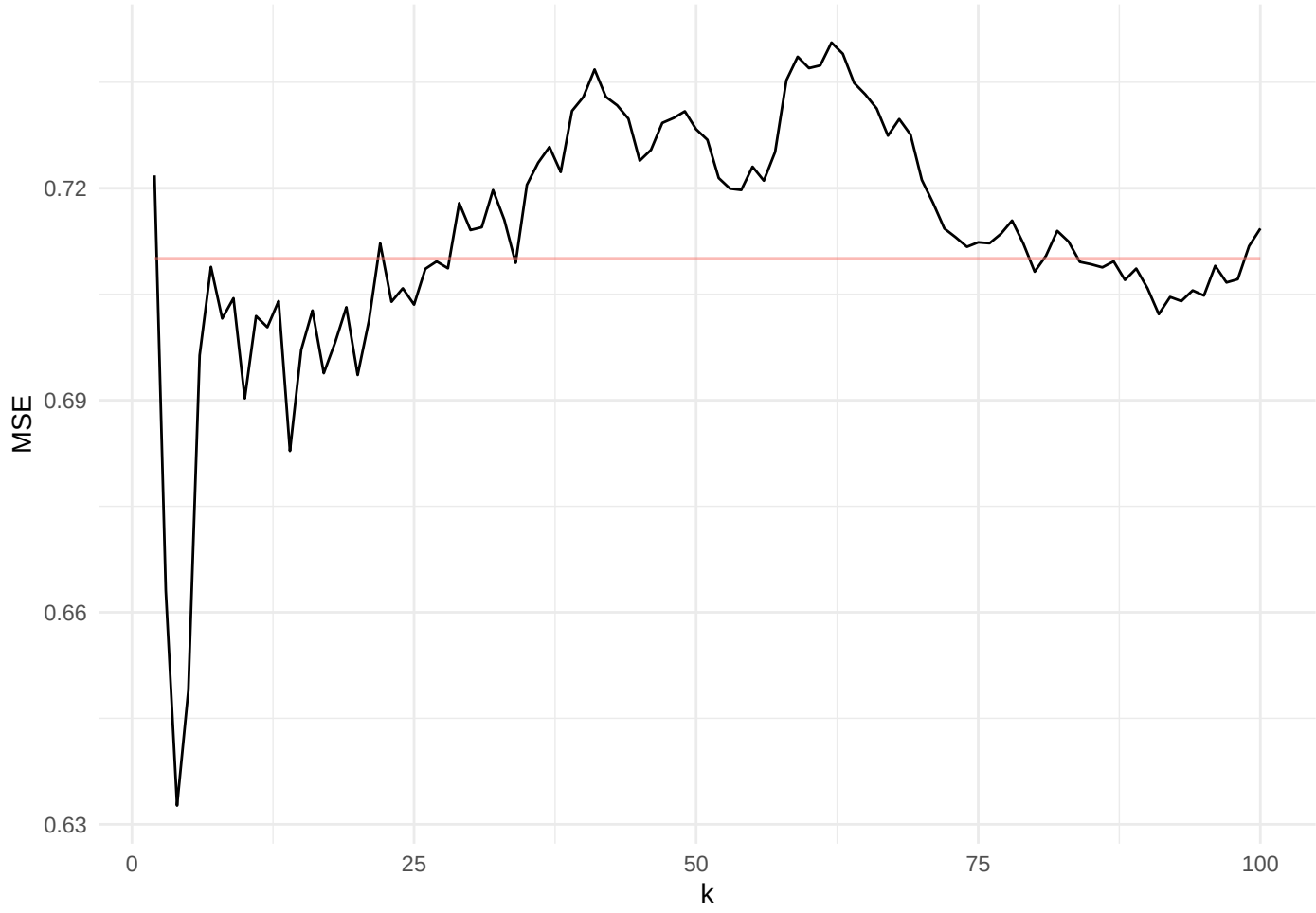
Best prediction by scenarios and Trait



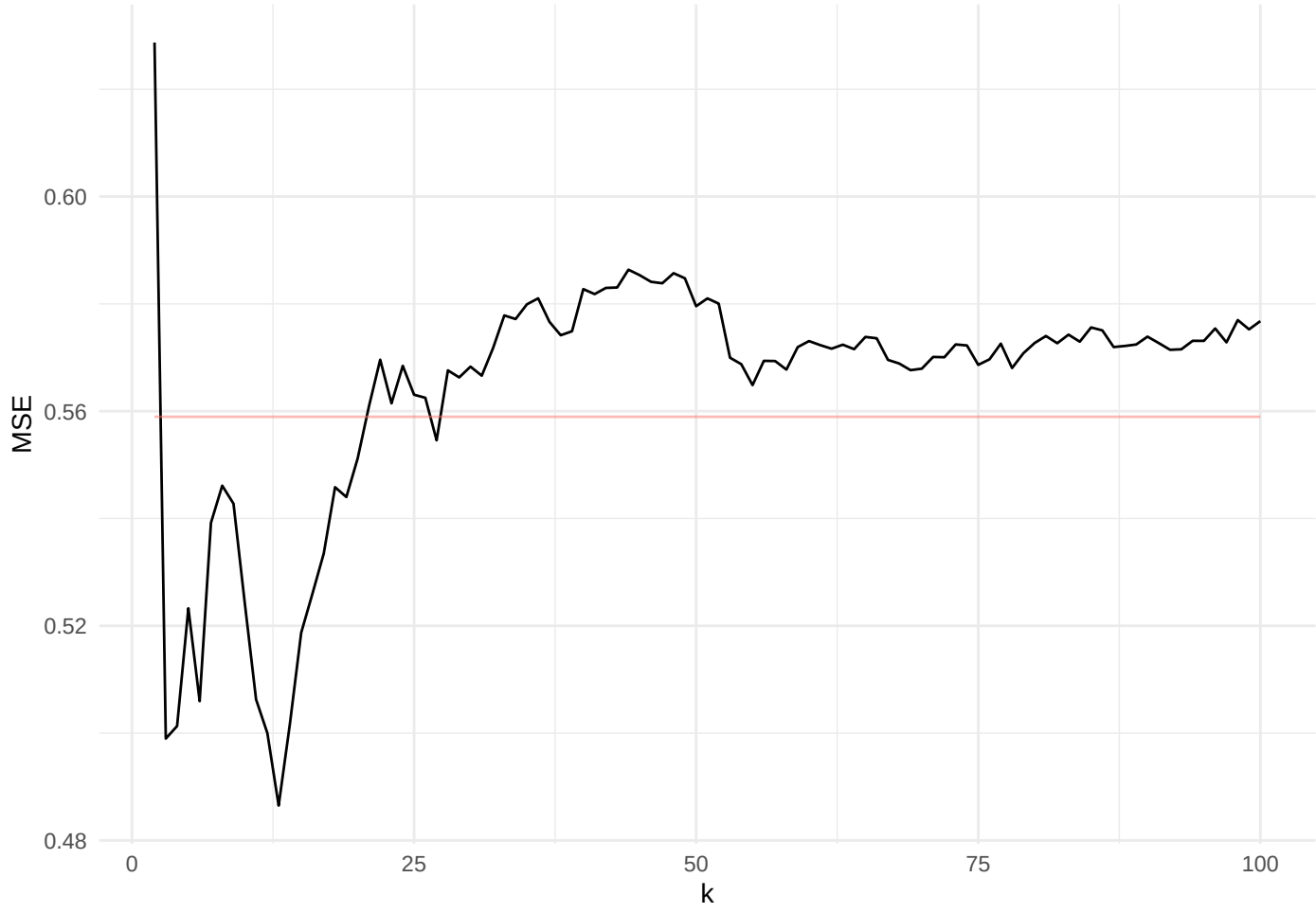
Neuroticism - Scenario 3



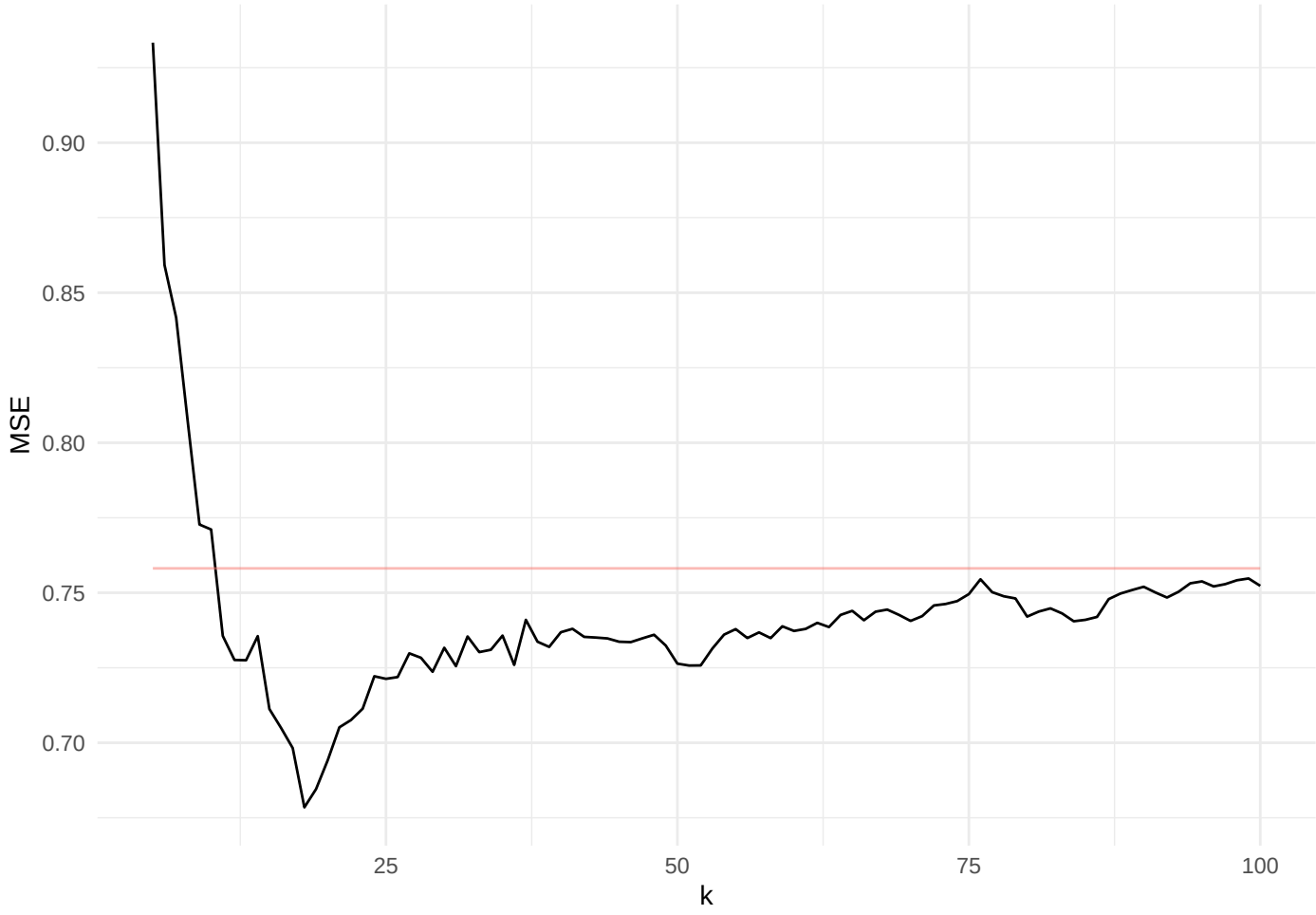
Neuroticism - Scenario 7



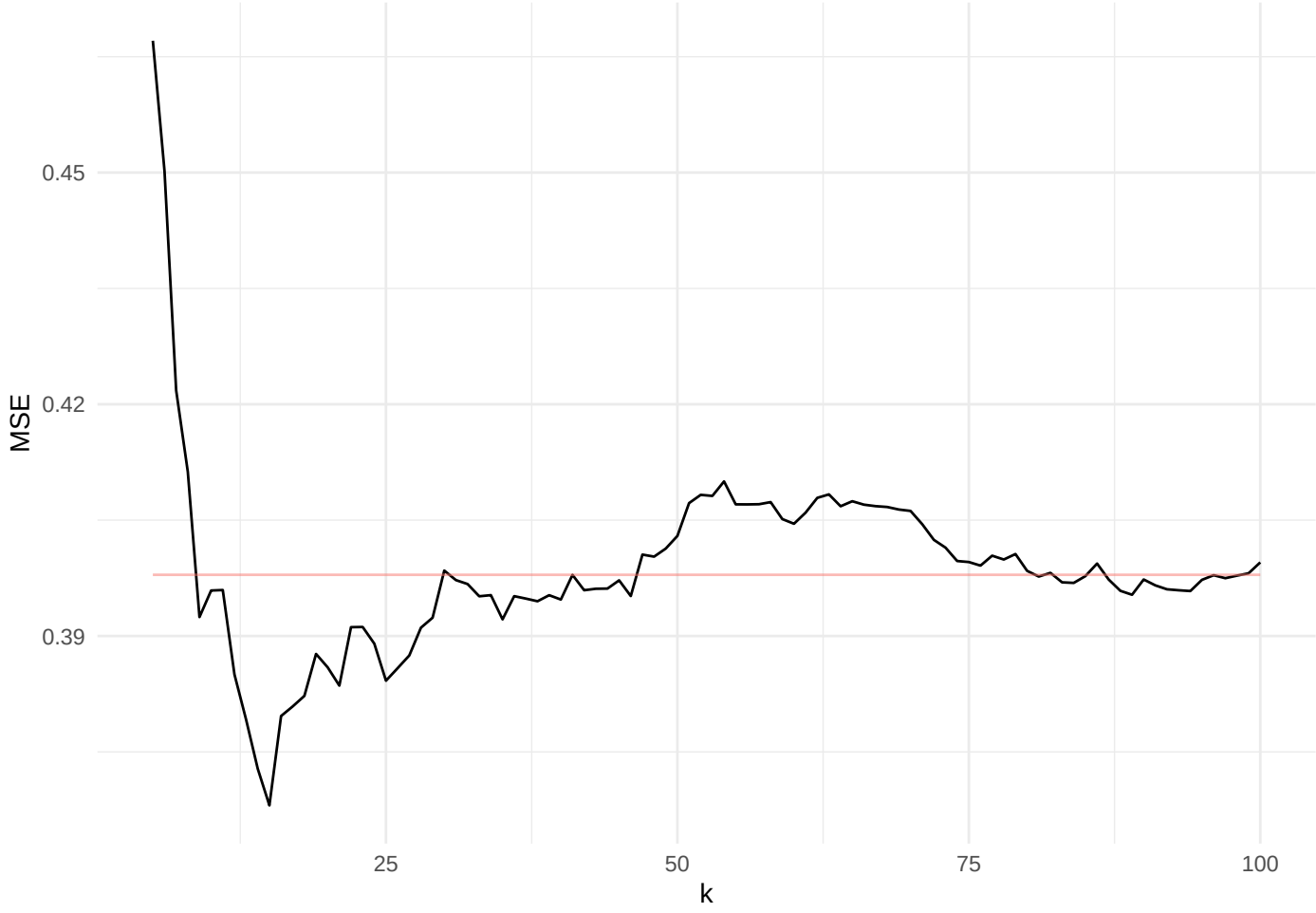
Openness



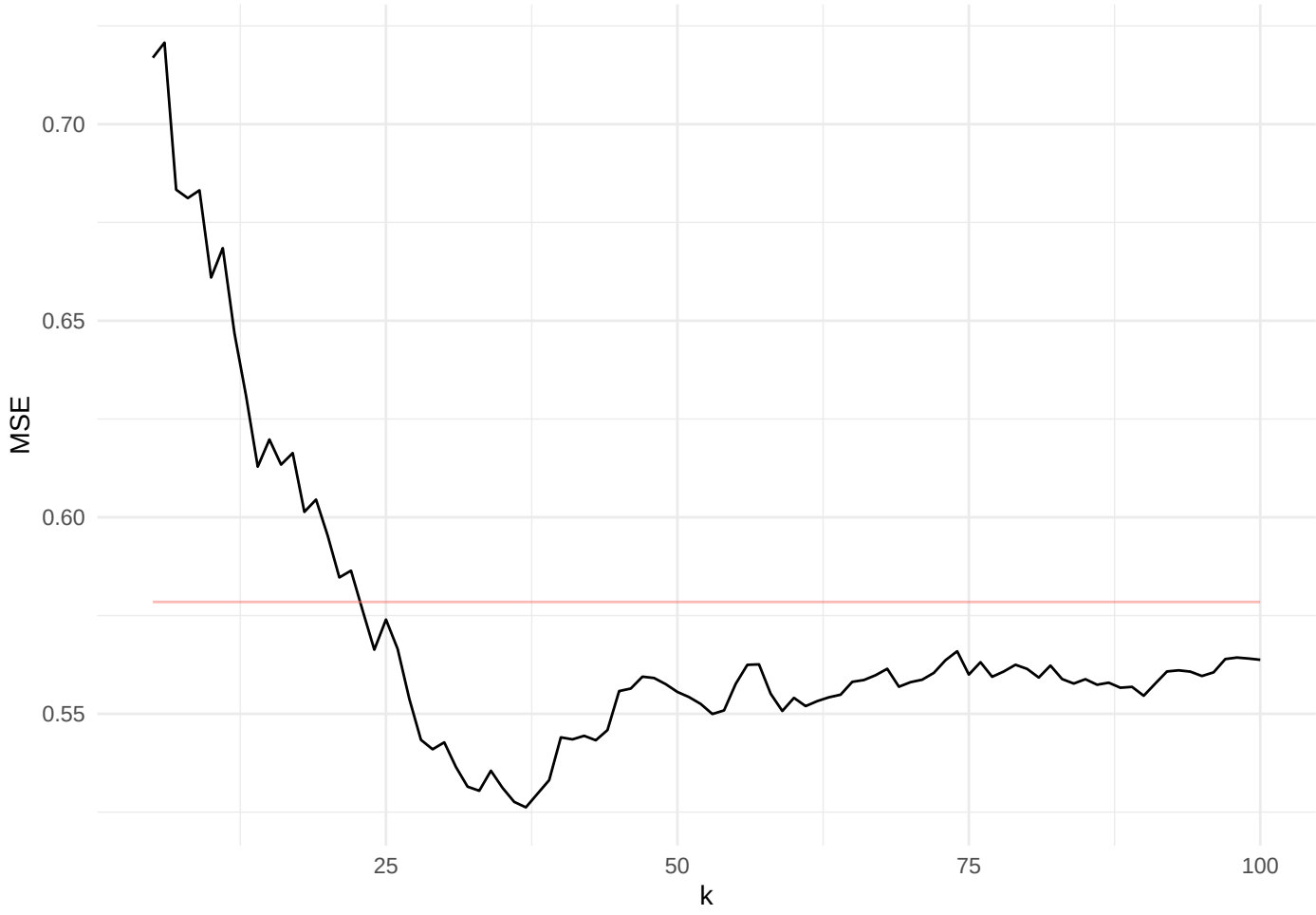
Extraversion



Conscientiousness



Agreeableness



Final comments

- OM provides a way to use categorical sequences for prediction.
- In most scenarios, the configuration of the clusters remained unchanged.
- The distance matrix highly sensitive to the definition of the cost matrix.
- Symmetry of the cost matrix may not be appropriate.
- The quality of the prediction is affected by the method chosen to handle missing values.
- The methodology makes it difficult to include additional variables for prediction.

References

- Optimal Matching Methods for Historical Sequences - A. Abbott & J. Forrest (1986)
- A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins - S. Needleman & C. Wunsch (1970)
- 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)