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

A Thesis

Presented to

The Division of Faculty of Science

University of Bern

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Statistics and Data Science

Adriana Clavijo Daza

June 2023

Approved for the Division
(Institute of Mathematical Statistics and Actuarial Science)

Prof. Dr. David Ginsbourger	Dr. Serena Lozza-Fiacco

Acknowledgements

I want to thank a few people.

Preface

This is an example of a thesis setup to use the reed thesis document class (for LaTeX) and the R bookdown package, in general.

Table of Contents

List of Tables

List of Figures

Abstract

The preface pretty much says it all.

Second paragraph of abstract starts here.

Dedication

You can have a dedication here if you wish.

Introduction

Optimal matching, introduced by Andrew Abbott Abbott & Forrest (1986) and since then has been extensively used in social studies to answer questions about processes that take values in a categorical space and occur along a period of time. Particularly, in sociology there have been several studies following the professional career of groups of people (e.g. graduates from a specific college or professionals in a certain area).

The main objective of these studies is to find similarities and differences between the sequences and obtain groups of trajectories that have common elements to them by means of a clustering method applied to the matrix distance obtained by optimal matching.

In this work, we explore the possibility to obtain predictions of a secondary variable based on the information obtained from the sequential data via optimal matching using the `R` package `TraMineR`.

We describe the optimal matching algorithm, normalization options and the methods for obtaining the cost matrix in Section 1. In Section 2, we introduce the data used in this work and perform optimal matching to obtain groups of sequences. In Section 3, we describe the method used for prediction with the data from the previous section. Finally, in section 4, we show some other scenarios that produce better predictions.

Chapter 1

Optimal Matching

Optimal Matching (OM) is a technique widely applied in social sciences for the comparison of sequences indexed by time that take values in a finite set of categories or states. For instance, in sociology, OM has been used for life course or career path analysis.

OM uses the Needleman-Wunsch algorithm, that was developed to compare biological sequences. This algorithm is an application of dynamic programming, an iterative method that simplifies an optimization problem by breaking it into a recursion of smaller problems that are simpler to solve.

The goal of OM is to find the best possible matching or alignment between two sequences by considering the differences and similarities between their elements and minimizing the total cost or dissimilarity between the sequences. The cost of comparing different states of the sequences can be defined in different ways including data-based methods or values supplied by experts in the particular field.

1.1 The OM algorithm

Consider a set of n categorical states $S = \{s_1, \ldots, s_n\}$, we define $X = (x_1, \ldots, x_t)$, a sequence of (discrete) length t, where $x_i \in S$ for $i = 1, \ldots, t$. Further, let $\mathbf{S}$ be the set of all possible sequences with states belonging to S.

Now, let $X, Y \in \mathbf{S}$ be two sequences of size t_X and t_Y, respectively. In order to numerically assess the similarities between the sequences X and Y, we define an empty array F of size $(t_X + 1) \times (t_Y + 1)$.

The algorithm below shows the initialization and recursion to fill the array F.

Here, the value d is the cost of inserting a gap in one of the sequences, also known

Algorithm 1 Optimal matching.

1: $F(1,1) \leftarrow 0$
2: **for** $j \leftarrow 2, t_Y + 1$ **do**
3: $\quad F(1,j) \leftarrow F(1,j-1) + d$
4: **end for**
5: **for** $i \leftarrow 2, t_X + 1$ **do**
6: $\quad F(i,1) \leftarrow F(i-1,1) + d$
7: **end for**
8: **for** $i \leftarrow 2, t_X + 1$ **do**
9: $\quad$ **for** $j \leftarrow 2, t_Y + 1$ **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**

as *indel* cost, and $K(y_{i-1}, x_{j-1})$ is the cost associated to change from the state y_{i-1} to x_{j-1}, which is defined in a matrix K of size $n \times n$, commonly known as the cost matrix.

Lines 1-7 of the OM algorithm correspond to initialization. Starting with a cost of 0 in $F(1,1)$, the first row and column of F represent cumulative costs of successively adding gaps. The remaining lines of the algorithm correspond to the row-wise recursion to fill the array F according to the content of the sequences to be compared: at any step of the recursion, the algorithm is looking at a specific pair of indexes (location) and calculating if substitution or insertion/deletion is the cheapest operation. Successively adding the costs of the cheapest operations results in the overall optimal cost for aligning (arrange to detect similarities) the sequences X and Y.

In fact, when F is completely filled, the value in the last cell, i.e. $F(t_X + 1, t_Y + 1)$ corresponds to the optimal cost of aligning the sequences X and Y. It is possible to recover the steps that conduced to this alignment with a traceback from the last cell. However, this is not necessary to obtain the dissimilarities matrix.

The `R` package `TraMineR` provides several functions to analyze and visualize sequential data. In particular, the package implements OM and offers several methods for computing the cost matrix K and the normalization of the dissimilarity matrix.

1.2 Cost matrix

The cost matrix K is a symmetric matrix of size $n \times n$. The value in the i-th row and j-th column $K(s_i, s_j)$ indicates the cost of moving from state s_i in time $t > 0$ to state s_j in $t + 1$.

The following are the methods available in `TraMineR`.

1.2.1 Transition rates (`TRATE`):

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

$$K(s_i, s_j) = c - P(s_i|s_j) - P(s_j|s_i), \tag{1.1}$$

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 $0 \leq K(s_i, s_j) \leq 2$).

1.2.2 Chi-squared distance (`FUTURE`):

$$K(s_i, s_j) = d_{\chi^2}(\mathbf{P_i}, \mathbf{P_j}), \tag{1.2}$$

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

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

$$K(s_i, s_j) = d_i + d_j, \tag{1.3}$$

where the *indel* cost d_i depends on the state and takes values:

$$g_i = \frac{1}{f_i}, \qquad \text{for method 'INDEL'}, \tag{1.4}$$

$$g_i = \log(\frac{2}{1+f_i}), \qquad \text{for method 'INDELSLOG'} \tag{1.5}$$

and f_i is the relative frequency of the state s_i for $i = 1, \ldots, n$.

Remarks:

- For methods `TRATE` and `FUTURE`, the unique *indel* value is $d = max_{1 \leq i,j \leq n} K(i,j)/2$, so that the cost of any transition is always lower or equal than deleting and inserting an element (or vice versa).
- The Needleman-Wunsch algorithm with constant costs for mismatch is known as Levenshtein distance.

1.2.4 Example

Let us suppose that S is the alphabet, let $X = \{S, E, N, D\}$ and $Y = \{A, N, D\}$ be two sequences in $\mathbf{S}$. Supposing that $d = 2$ and

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

The array F is initialized as follows:

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

To fill the second row of F we proceed as follows:

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

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

Finally, we obtain the following F array:

		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

In this simple example, we can easily obtain two optimal (equivalent) alignments without using the algorithm.

S E N D with

A – N D or

– A N D

In both cases we have two matches (cost 0), one mismatch (cost 3) and one gap (cost 2), giving a total cost 5 that is exactly what we obtained in the last cell of F.

The cost of inserting a gap (d) is also known as *indel* (insert or delete) cost. In this example we can observe that, in order to obtain sequence X from Y we have to **insert** a term (i.e. insert a gap and then change its value to a specific state). Equivalently, to obtain sequence Y starting from X we have to **delete** one term.

1.3 Normalization

In cases when the lengths of the sequences differ, it can be useful to account for this differences with a normalization factor.

Given a set two sequences $X, Y \in \mathbf{S}$ 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` offers the following options to normalize the distances between sequences:

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

Chapter 2

Data from the 40+ Healthy Aging Study

2.1 About the data

As part of the Women 40+ Healthy Aging Study, a large study that was conducted by the Department of Clinical Psychology and Psychotherapy of the University of Zurich, a psychometric instrument was developed in order to obtain information about the history of romantic relationships of women. The study was conducted between June 2017 and February 2018 with women between 40 and 75 years who (self-)reported good, very good or excellent health condition and the absence of acute or chronic somatic disease or mental disorder. The participants who reported psychotherapy or psychopharmacological treatment in the previous 6 months were excluded as well as habitual drinkers. Other exclusion criteria were pregnancy in the last 6 months, premature menopause, surgical menopause, intake of hormonal treatment (including contraceptives), shift-work and recent long-distance flight. The participants were recruited from the general population using online advertisement and flyers.

The questionnaire asked the participants to provide information about relationship phases starting from the age of 15 years until the current age at the time of the data collection. The phases were 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. Before including the data corresponding to their own history, the participants were prompted to answer some of the questions based on an example. Some of the participants were excluded when the example entries were not correctly filled. After data cleaning and revisions for consistency the

total number of individuals considered is 239.

In order to create a sequence for each participant the information about civil status, relationship status, living situation and the maternity is taken into account. A yearly sequence is created and the states considered are the following:

- 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

Additionally, personality scores for the women included in the study are available. Personality refers to the enduring characteristics and behavior that comprise the unique adjustment to life of a person, including major traits, interests, drives, values, self-concept, abilities, and emotional patterns. These scores are obtained via psychometric instruments and evaluate the main personality traits:

- Agreeableness
- Conscientiousness
- Extraversion
- Neuroticism
- Openness

2.2 Application of OM

Using the `R` package `TraMineR` the cost matrix is calculated with transition rates between states. We consider a base setup with method `TRATE` for the calculation of the cost matrix and `maxlenght` normalization for the dissimilarities matrix. The obtained cost matrix is shown below. As expected, the elements in the diagonal are equal to 0, meaning there is no cost associated to staying in the same state. By default, the constant c in 1.1 is set to 2. This, and the fact that the duration of the states is often longer that the time unit (one year), makes that all of the values outside the diagonal are close to 2 and even equal in cases where no transition between the states were

Table 2.1: Cost matrix obtained from transition probabilities.

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

observed in the data (e.g. from single without children to single with children and vice versa). Finally, we observe that missing value (`NA`) is considered as a separate state and, by default, the cost of changing from or to a missing value is 2, which might be too high in cases where the individuals made a mistake in the beginning or end age of a phase leaving a gap in the sequence.

From this cost matrix it is possible to calculate pairwise distances between all the sequences using the algorithm described in the previous section. As stated before, a correction of the distances is done to account for the differences in size of the sequences. This is done dividing the obtained distance by the length of the longest sequence.

Having obtained the distance matrix, we apply a hierarchical agglomerative clustering method in order to explore the data and the differences captured by the distance matrix. In particular, we set the number of clusters to 4 and the following figure shows the distribution of the states.

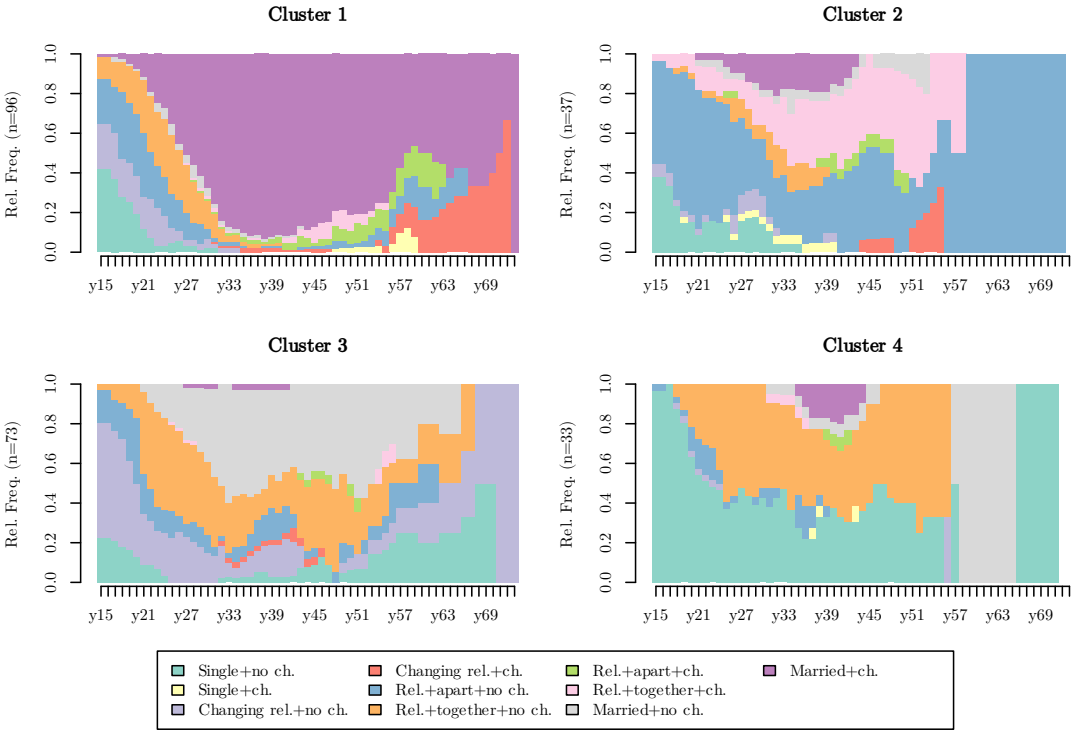

Figure 2.1: Distribution of states by cluster.

The figure below shows the transverse entropy by cluster, i.e. the cross-sectional entropy of the states distributions is calculated at each time point as follows:

$$h(f_1, \ldots, f_n) = -\sum_{i=1}^{n} f_i \log(f_i). \tag{2.1}$$

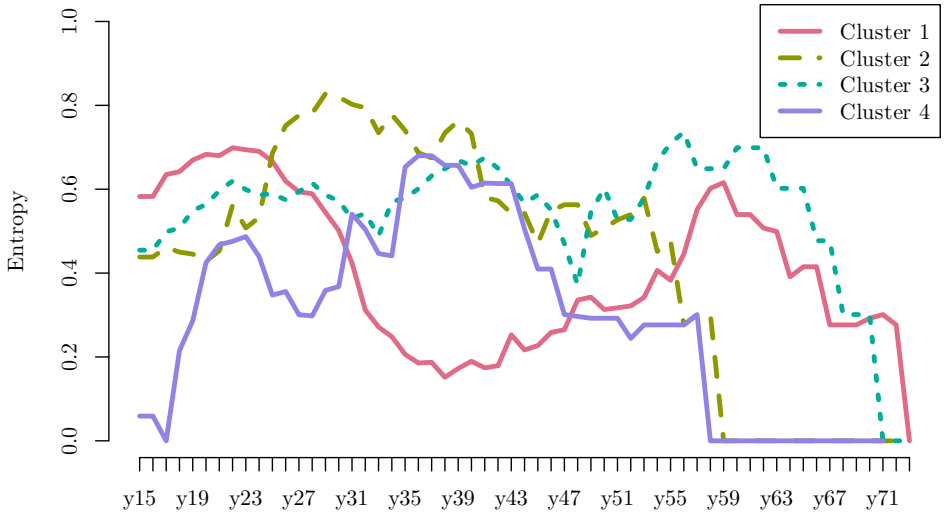

Figure 2.2: Transversal entropy by cluster.

Table 2.2: Number of individuals by cluster.

Cluster	n
1	96
2	37
3	73
4	33

The previous visualizations allow us to try to identify common and contrasting features of the clusters that can be useful to describe them. It is important to remember that this description is subjective and incomplete.

- 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.

On the other hand, in figure 2.1 we can also appreciate that the conformation of some clusters seems to be highly affected by the length of the sequence and is possible that the normalization method is not achieving the expected result.

We are interested in exploring how the relationships history of the women relate to personality traits. As a first exploratory step, the following figure shows the distribution of the score for each trait by cluster.

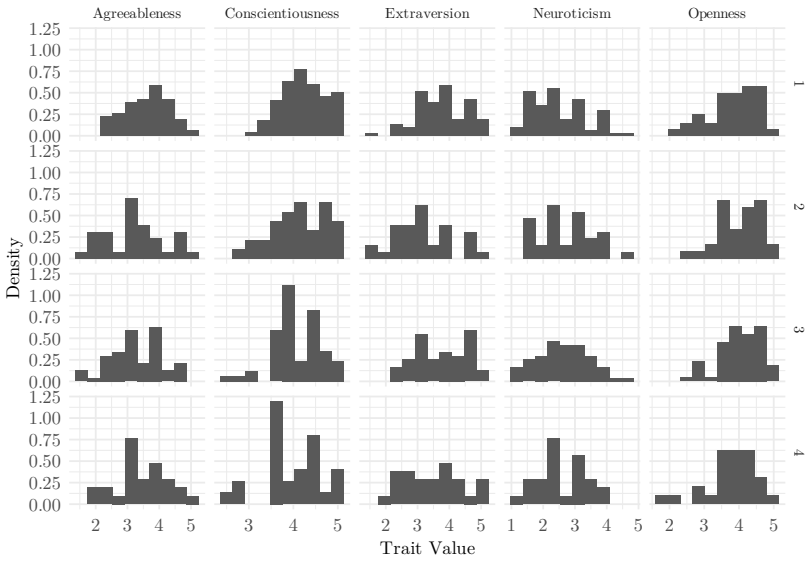

Figure 2.3: Distribution of personality scores by cluster.

No difference is obvious at first glance. Also, the number of clusters and the

fact that the personality scores are not continuous makes it difficult to appreciate differences. For that reason, we also explore with a lower number of clusters.

Furthermore, we obtain better defined clusters that are less affected by the length of the sequences as we can observe in the distribution plots of the sequences states: the majority of women in cluster 1 have children, while we mostly find women without children in cluster 2.

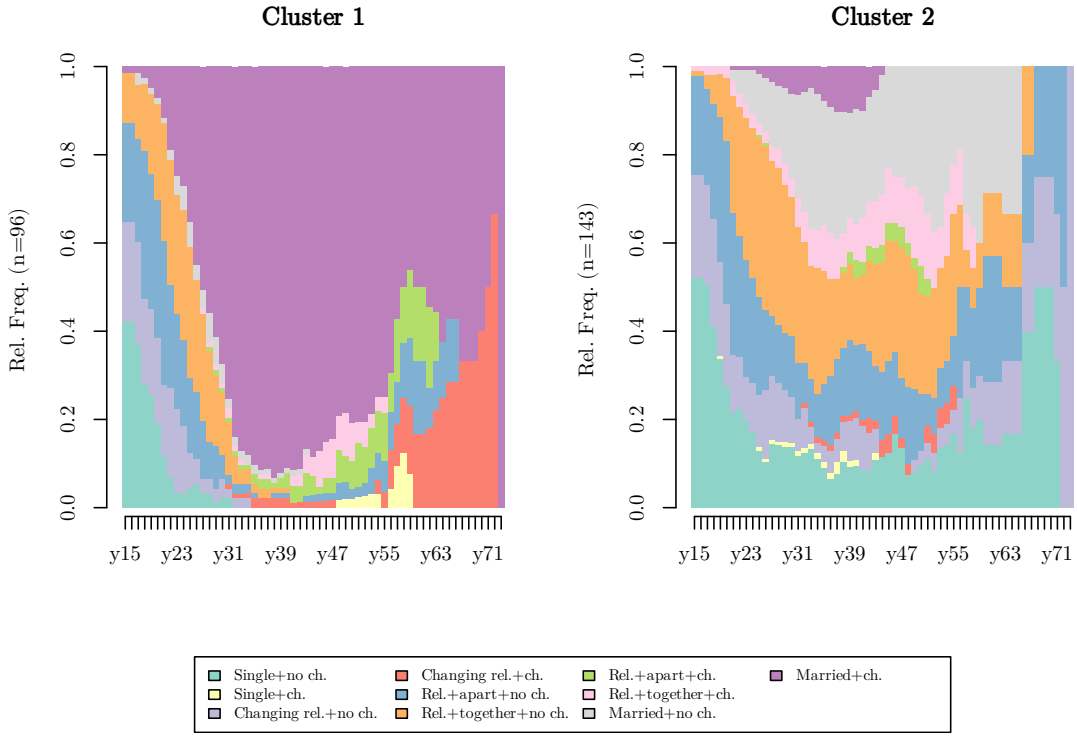

Figure 2.4: Distribution of states for two clusters.

In addition, the transversal entropy of the sequences for the two clusters is displayed in the figure below.

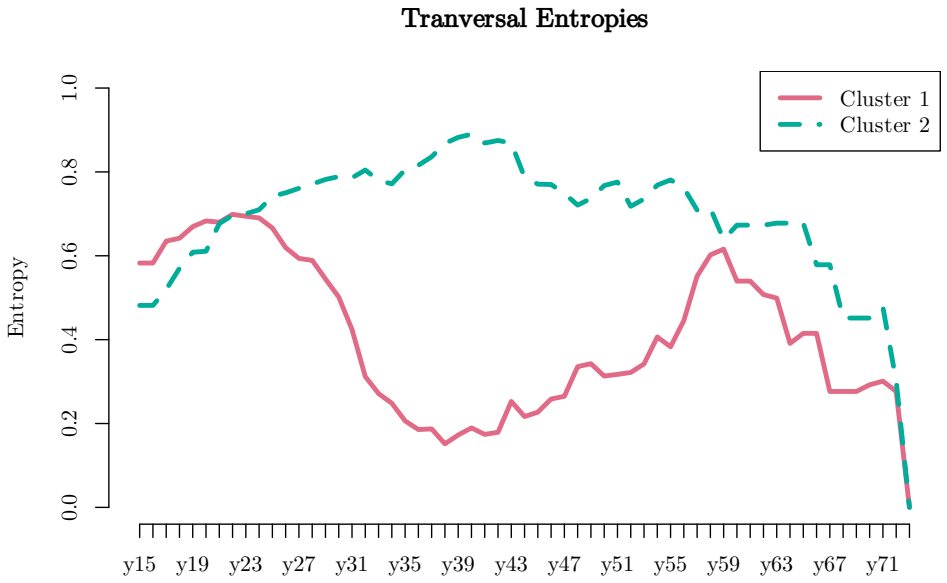

Figure 2.5: Transversal entropy for two clusters.

This figure shows that the entropy decreases significantly around mid age for the cluster of women with children as compared to women without children, which means that the variability of the states for the first group is much lower as compared to the second group. This can be interpreted as a sign of stability in the relationship status for women during the time they have children at home.

As before, we want to explore possible links between the information from the sequences and personality scores. The following figure shows the distribution of the personality traits for the two clusters.

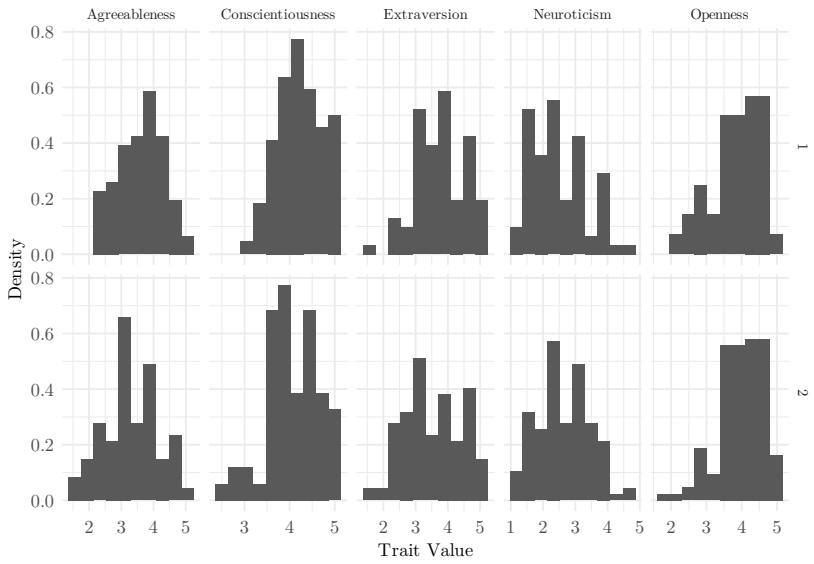

Figure 2.6: Distribution of personality scores for two clusters.

There seems to be differences in the distributions of some personality scores: the scores of agreeableness are concentrated in larger values for women with children; women without children have greater frequency in lower values of conscientiousness than women with children; and women with children exhibit lower scores of neuroticism.

Even though, the distribution of personality scores by cluster does not reveal significant differences, the obtention of a distance matrix also provide us a numerical expression of the categorical sequences that allows us to use it for other purposes. In particular, we explore the predictive capability of this data with a non-parametric prediction method in the next section.

Chapter 3

Personality Scores Prediction with k-Nearest Neighbors

Given a 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 class labels for classification or a continuous range of values for regression). k-NN provides a way to predict the label or value for a new, data point x_{n+1} (for which Y is unknown) 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).

There are different ways of calculating the distance between the new data point x_{n+1} and the points in $\mathcal{D}$. For instance, the Euclidean o Mahalanobis distances are usually used. In our case we already count with a matrix distance obtained with OM.

The choice of k is a hyperparameter that can be tuned to optimize the performance of the k-NN algorithm. A larger k reduces the effect of noise and outliers, but can also lead to overfitting. A smaller k is more sensitive to noise and outliers, but can better capture local structure.

To compare the performance of different values of k, we use the mean squared error (MSE).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{3.1}$$

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

In this part of the analysis we only consider the individuals who have available personality scores, that leaves us with a sample size of 200 individuals. We also split the data into two subsets: train (70%) and test (30%). We evaluate the MSE of the predictions for the individuals in the test set but only using the data from the nearest neighbors available in the train set.

The following figure shows for every personality trait and different values of k the MSE, i.e. for $k = 1, \ldots, 80$ we predict values of Y and compare them with the observed values using the MSE. As a reference, a red line for every personality trait is added to indicate the MSE of the trivial prediction, i.e. the prediction considering all the sample points in the train set.

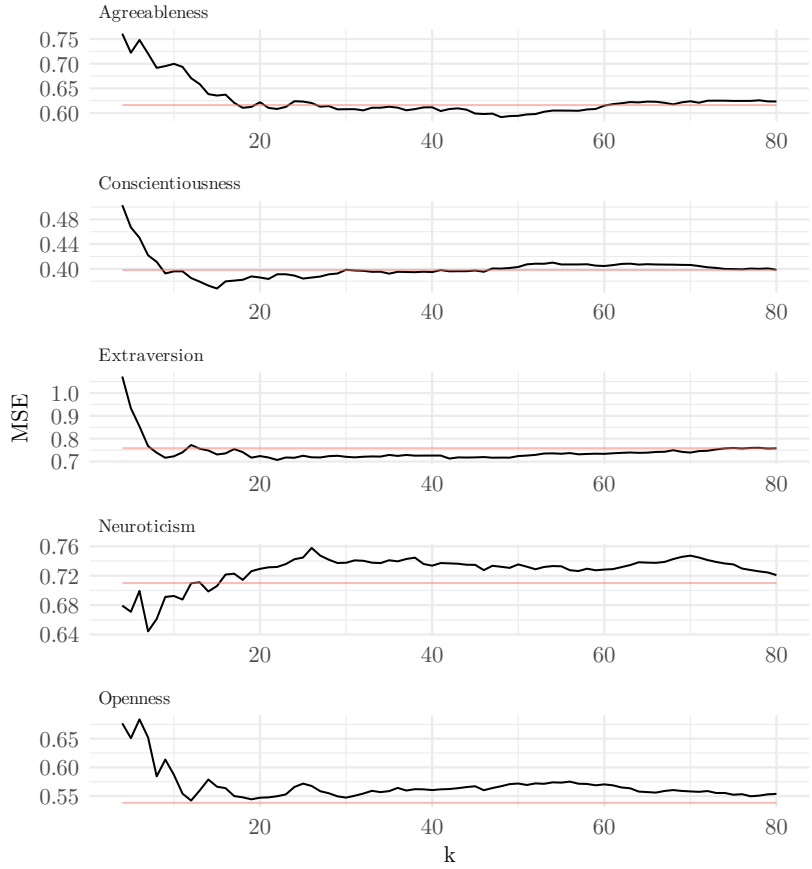

Figure 3.1: MSE by personality trait for base setup prediction.

Overall, it seems that using the sequential data for prediction results in little improvement compared to the trivial prediction except for neuroticism where the MSE takes a minimum value around $k = 5$.

Furthermore, for conscientiousness and openness, the MSE does not seem to increase again as k increases, which is expected when using kNN, due to overfitting. Moreover, for openness, the prediction with kNN is always worse than the trivial prediction. For conscientiousness, the MSE takes a minimum value around $k = 15$ and after $k = 30$ the MSE curve stays flat.

For agreeableness, the MSE is minimum around $k = 50$ and increases again. However, this minimum is not considerably lower than the trivial prediction. Similarly,

for extraversion, the MSE takes a minimum value after $k = 20$, but is not a significant improvement compared to the trivial prediction.

Given that the performance of the predictions is just slightly better than average in most cases, we contemplate other scenarios with different variations of the hyperparameters considered in this section.

Chapter 4

Additional scenarios considered for prediction

In order to find better prediction for the personality scores, we considered different configurations for the obtention of the cost matrix. For instance:

- Take the constant c in 1.1 as $2 * max_{1 \leq i,j \leq n} P(i,j)$ so that $0 \leq K(s_i, s_j) \leq 2 * max_{1 \leq i,j \leq n} P(i,j) \leq 2$).
- Consider the methods `FUTURE`, `INDELS` and `INDELSLOG` for the calculation of the cost matrix.
- Try `gmean` and `maxdist` as the normalization factor for the distance matrix.
- Consider several values from 0 to 2 for the transition from/to missing value, given that this has a significant effect when comparing sequences with large differences in length. Also, we can appreciate this effect in the conformation of the clusters (see 2.1).
- Given that the previous consideration resulted in better prediction performance, and with the aim of obtaining more homogeneous sequences in length, we limit the start and end age of the sequences.

The following table shows some of the scenarios considered. With the purpose of comparing the predictions obtained with the different scenarios, we calculate the relative improvement (p) compared to the trivial prediction for each value of k and each scenario.

$$p = (1 - (MSE_k / MSE_{trivial})) * 100 \tag{4.1}$$

Table 4.1: Summary of the additional scenarios considered

	Cost matrix	Normalization	Transition constant	NA cost	Min age	Max age
1	TRATE	maxlength	NULL	NULL	NULL	NULL
2	TRATE	maxlength	0.08021433	NULL	NULL	NULL
3	TRATE	gmean	0.08021433	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	maxdist	NULL	NULL	NULL	NULL
9	FUTURE	gmean	NULL	1	NULL	NULL
10	FUTURE	gmean	NULL	NULL	20	55
11	FUTURE	gmean	NULL	NULL	20	40
12	FUTURE	gmean	NULL	1	20	55
13	FUTURE	gmean	NULL	0.5	20	55

The following figure shows the best relative improvement achieved for each personality trait and under all the scenarios in 4.1 and the corresponding value of k at which the best performance was obtained.

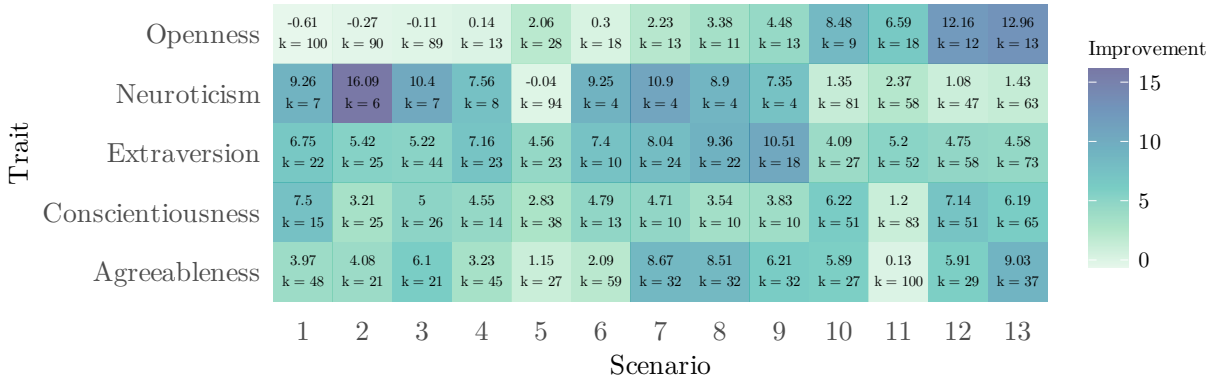

Figure 4.1: Relative MSE improvement in the prediction of personality traits.

We can observe that there is not a single scenario that produces the best prediction improvement for every trait. Although, the method `FUTURE` seems to produce better results for all of the traits except neuroticism.

The following figure shows the MSE for neuroticism in scenario 2 in which the cost matrix is calculated with transition rates, the normalization method for the distances is `maxlength` and the constant c was modified.

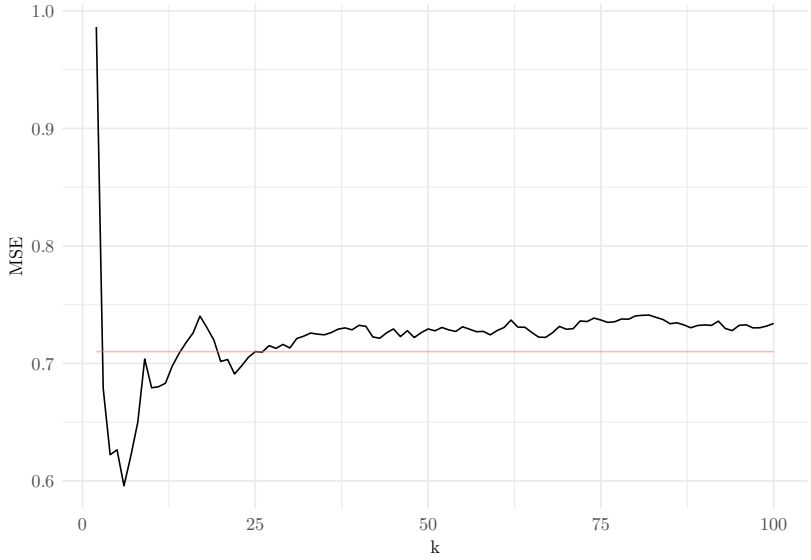

Figure 4.2: MSE of neuroticism prediction in scenario 2.

We can observe that the best prediction is achieved at $k = 6$. However, the MSE at this point is much lower than the rest of the curve. It might be the case that this is a random occurrence. However, in 4.1 we observe the minimum MSE with similar number of neighbors for other scenarios, for instance, in scenario 3.

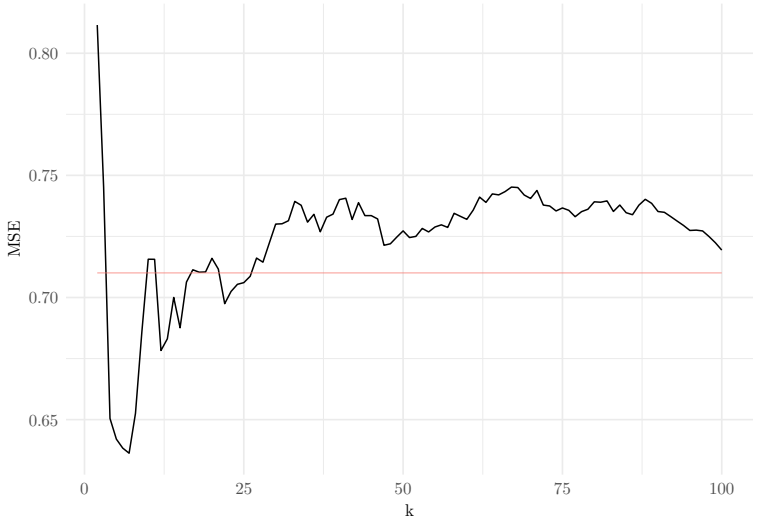

Figure 4.3: MSE of neuroticism prediction in scenario 3.

For openness, the best predictions are obtained with scenarios 12 and 13. Both scenarios consider χ^2 distances for the calculation of the cost matrix (method `FUTURE`), normalization with the method `gmean` and the sequences are restricted between 20 and 55 years of age. This scenarios differ in the cost assigned to changes involving missing values. Hence, we might infer that the prediction of openness is highly affected by the

way missing values are handled. The following figure shows the MSE for openness in scenario 13.

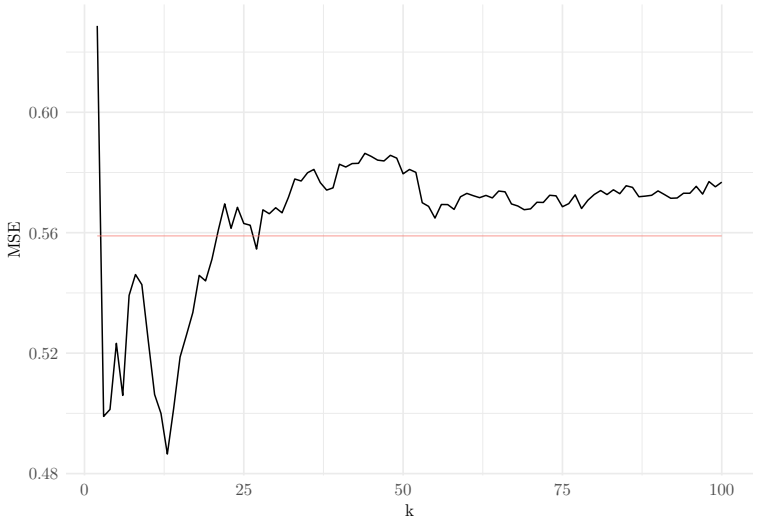

Figure 4.4: MSE of openness prediction in scenario 13.

In this case, we observe that the curve is lower around the values near to where the minimum is obtained at $k = 13$ and it increases to values where the performance is worst than the trivial prediction from $k > 25$.

The scenario that produces the best prediction for extraversion is number 9. In the figure below the MSE for this scenario is shown.

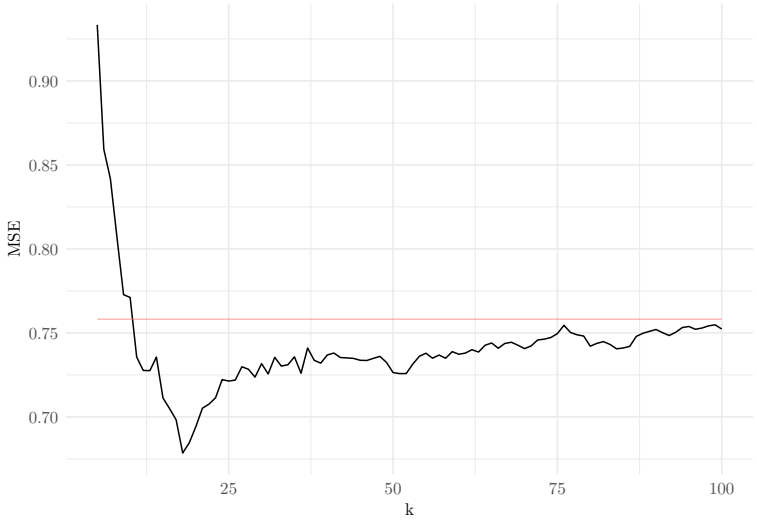

Figure 4.5: MSE of extraversion prediction in scenario 9.

In this case, the minimum MSE is obtained when $k = 18$ which is a high number of neighbors compared to the two previous traits. As expected, the MSE decreases

until this value and then starts to increase again, a sign of overfitting for bigger values of k.

For conscientiousness none of the predictions achieved a relative improvement of at least 10%. Furthermore, the best prediction is obtained for $k = 15$ in the base scenario and also in scenario 2 (see following figure) where the $c = 2 * max_{1 \leq i,j \leq n} P(i, j)$, which is consistent with the fact that changing the value of c implies a translation of the distances but does not affect their ordering of the neighbors.

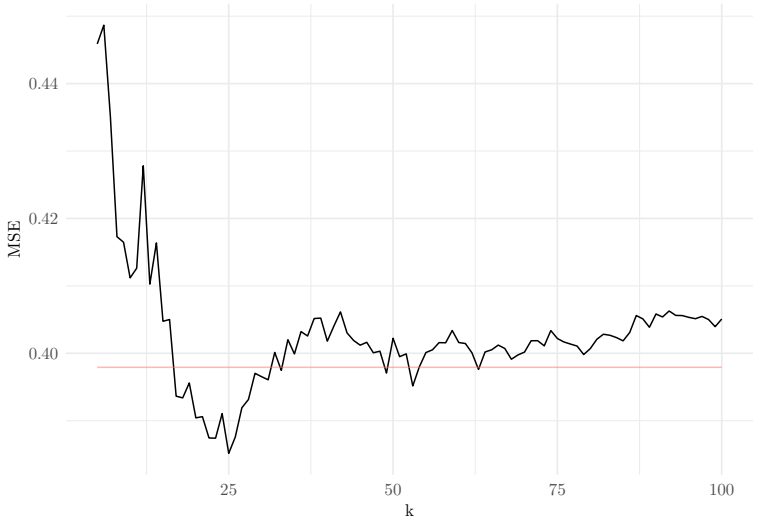

Figure 4.6: MSE of conscientiousness prediction in scenario 2.

Likewise, for agreeableness, all of the scenarios showed improvements relative to the trivial prediction and the minimum MSE in every case is obtained for large values of k. This could be an indication of poor predictive power of the relationships history of women for this particular trait. However, as expected for this prediction technique, we observe in the following figure that the MSE is large and even greater than the MSE of the trivial prediction for values of k below 25 and after achieving the minimum it starts increasing again.

Table 4.2: Cost matrix for scenario 13.

State	1	2	3	4	5	6	7	8	9	10	NA
1	0.00	1.31	1.24	1.28	1.18	1.18	1.25	1.26	1.28	1.25	0.5
2	1.31	0.00	1.30	1.33	1.28	1.28	1.27	1.32	1.34	1.33	0.5
3	1.24	1.30	0.00	1.27	1.17	1.14	1.24	1.26	1.26	1.23	0.5
4	1.28	1.33	1.27	0.00	1.25	1.24	1.22	1.24	1.31	1.29	0.5
5	1.18	1.28	1.17	1.25	0.00	1.15	1.20	1.24	1.24	1.22	0.5
6	1.18	1.28	1.14	1.24	1.15	0.00	1.21	1.23	1.23	1.20	0.5
7	1.25	1.27	1.24	1.22	1.20	1.21	0.00	1.24	1.28	1.26	0.5
8	1.26	1.32	1.26	1.24	1.24	1.23	1.24	0.00	1.31	1.28	0.5
9	1.28	1.34	1.26	1.31	1.24	1.23	1.28	1.31	0.00	1.30	0.5
10	1.25	1.33	1.23	1.29	1.22	1.20	1.26	1.28	1.30	0.00	0.5
NA	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.0

Figure 4.7: MSE of agreeableness prediction in scenario 13.

Finally, we perform clustering again with the distance matrix obtained in Scenario 13. The following table shows the cost matrix for this setup of parameters. We can appreciate that in this case the range of the values of the cost matrix is larger than those observe in 2.1. The figure below show the distribution of states by cluster for this scenario.

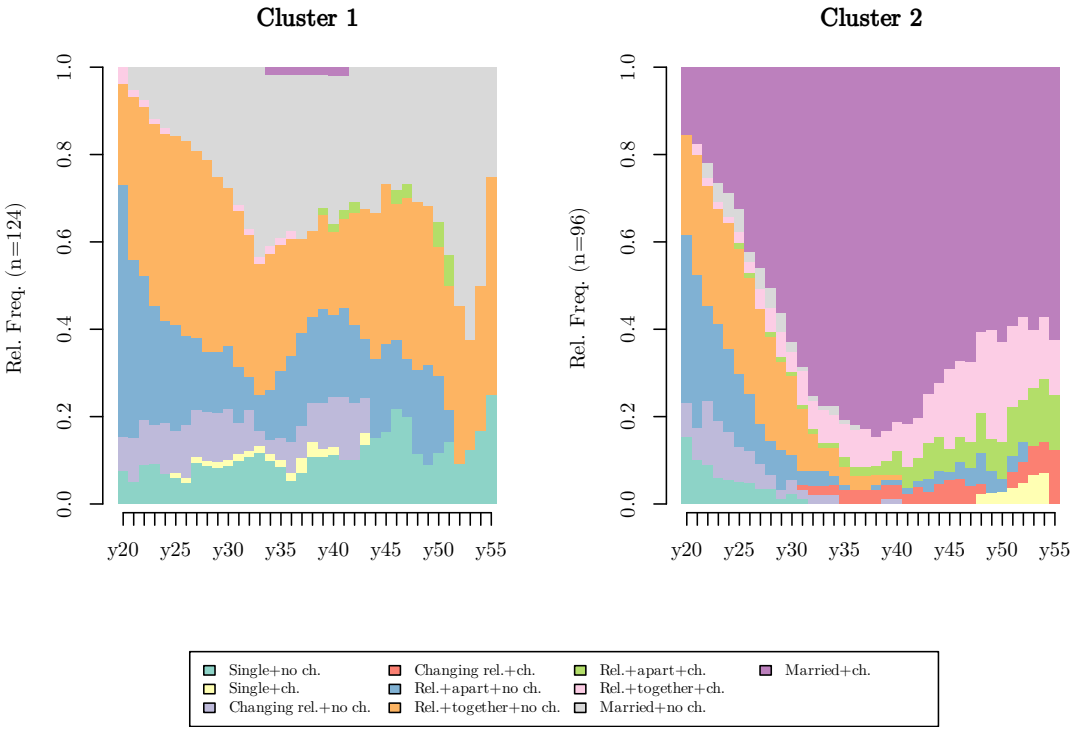

Figure 4.8: Distribution of states for two clusters in Scenario 13.

Even tough the cost matrix presented large variations compared to the base scenario and the predictions improved, we obtain similar clusters: In cluster 1, we find women with different relationship situations but without children. Similarly, in cluster 2 we find women with different trajectories that eventually had children.

References

Abbott, A., & Forrest, J. (1986). Optimal matching methods for historical sequences. *The Journal of Interdisciplinary History*, *16*(3), 471–494. Retrieved from `http://www.jstor.org/stable/204500`

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. http://doi.org/`https://doi.org/10.1016/0022-2836(70)90057-4`