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.

Table of Contents

List of Tables and Figures

Abstract

In this work, we describe and use optimal matching (OM) — a method widely used in sociological studies — to obtain pairwise distances between categorical sequences that represent the relationship and family history of a group of women over 40 years old. We then use the distances to obtain clusters and predict five personality traits with k-nearest neighbors (k-NN) and compare the performance of different approaches using the mean squared error (MSE) of the prediction. The different approaches or scenarios are given by the calculation of the cost matrix, normalization and handling of missing values in the sequences. Our aim is to explore the effects that the variations considered in the scenarios have on an unsupervised and supervised distance-based method.

Dedication

You can have a dedication here if you wish.

Introduction

In order to extract useful information from a dataset of categorical sequences, we can obtain pair-wise distances between the data points and use a distance-based method which is particularly useful when dealing with complex data types. Depending on the objective of the researcher and the availability of other observed variables, we can apply an unsupervised or supervised learning technique. For example, we can obtain groups of similar sequences, identify common trajectories or predict other variables of interst.

In computational linguistics, we encounter several methods to obtain these distances, for instance, there is a category of measures known as *edit distances* that are widely used for text prediction. To give some examples, the Hamming distance accounts for differences at each time point or we can define a dissimilarity based on the length of common subsequences or prefixes. One of the measures considered as an edit distance is the Levenshtein distance which finds the minimum number of steps required to arrive to one sequence taking the other as a starting point. The steps can be insertion, deletions or substitutions in the case of the Levenshtein distance, but other distances also consider the option of character swaps.

In bioinformatics, a generalization of the Levenshtein distance was proposed by Needleman & Wunsch (1970) to find similarities in the amino acid sequences of two proteins. The Needleman-Wunsch algorithm finds the best alignment of two sequences by maximizing the similarity between them with the possibility of different penalization values for substitutions. This algorithm was introduction and adapted to social sciences by Abbott & Forrest (1986), who named it optimal matching (OM). Since then, it has been extensively used to answer questions involving sociological processes that take values in a categorical set and occur along a specific period of time, mainly in the study of paths of family formation or professional careers.

The common result of several studies employing OM is to find similarities and dissimilarities among categorical sequences and identify groups of trajectories that exhibit similar patterns. This is achieved through the application of clustering

techniques to the distance matrix generated with help of the algorithm. However, there are a number of decisions involved in the application of OM, namely the method for calculation of the *cost matrix* or the normalization applied to the distances. We can see these decisions as hyperparameters that need to be tuned.

In this work, we use a dataset that contains categorical sequences of the relationship history from the age of 15 years to the current age at the time of data collection for a group of women in the context. The sequences contain information about the civil status, relationship status, cohabitation status and presence or absence of children. The data was obtained as part of the 'Women 40+ Healthy Aging Study' which also encloses scores of personality traits that were obtained with a psychometric instrument.

We explore then, the effect that changing the hyperparameters has on the distance matrix obtained via OM by considering how groups obtained via hierarchical clustering change and how the quality of the predictions of the personality scores is affected.

For this purpose, we use the implementation of the OM algorithm in the `R` package `TraMineR` (Gabadinho, Ritschard, Müller, & Studer (2011)) and the visualizations provided by `TraMineRextras`.

In Section **??** we give an account of the most commonly used edit distances, then we introduce the data we use in out application, explain the method chosen for prediction and provide a detailed description of the optimal matching algorithm, including the methods considered to obtain the cost matrix and the normalization methods for the distance matrix. In Section 2, we provide more insight on the data and perform OM to obtain clusters of sequences and use visualizations to provide a description of the groups. In Section 3, we describe perform a first attempt at prediction of the personality scores based on the data from the previous section and in Subsection 3.3, we show some of the additional scenarios considered, specifically those that produce the best predictions for each personality trait. Finally, we present some conclusions and recommendations for future work.

Section 1

Distance-based methods for categorical sequences

Distance-based methods are a class of statistical techniques based on the use of distance, similarity or dissimilarity between data points defined by a distance or similarity function. The main idea behind these methods is to obtain a (pseudo)distance matrix in order to apply an unsupervised learning method such as

- clustering in different variations
- dimensionality reduction
- multidimensional scaling

as well as supervised learning methods, where we are interested in using the data to predict other variable for which we have observed or labeled values, for example: k-nearest neighbors (k-NN).

Distance-based methods allow application with a variety of data types, in particular, categorical sequences, i.e. sequences that take values in a finite set of categories or states and are indexed by time.

1.1 Distance in categorical sequences

Several ways to compute distances between categorical sequences have been proposed in the context of natural language processing and bioinformatics. Particularly, a class of measures, known as *edit distances*, provide a quantification of the dissimilarity of a pair of sequences by counting the minimum number of operations required to obtain a sequence from the other. For instance, Hamming (1950) proposed a distance for sequences of the same length that counts the number of positions with different

states. Levenshtein (1966) generalized the Hamming distance to sequences of different lengths by considering the minimum number of single character edits required, namely insertion, deletion and substitution. The Damerau-Levenshtein distance allows for an additional operation known as transposition or swaping of characters (Damerau (1964)). The Jaro-Winkler similarity measure counts the number of matches and transpositions but does not fulfill the triangle inequality (Winkler (1990)). Additionally, there are several algorithms to find the longest common subsequence from a pair of subsequences (see Bergroth, Hakonen, & Raita (2000)), this could be seen as a measure that allows for insertion and deletion but not substitution nor transposition.

Also, note that from fundamental statistical methods for categorical data, considering the distribution of the states in a sequence, for a pair of sequences, the euclidean or χ^2 distance can be obtained and used in this context.

1.2 Optimal matching and applications

One generalization of the Levenshtein distance that allows for different substitution penalties for every pair of states is the Needleman-Wunsh algorithm, developed by Needleman & Wunsch (1970) with the aim of comparing biological sequences (for example, DNA or protein 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. By choosing the optimal operation at each step, it is guaranteed that the overall solution is optimal as well. An adaptation of this algorithm, known as *optimal matching* (OM), was introduced in social sciences by Abbott & Forrest (1986) and has been widely applied in sociology, for instance, optimal matching has been employed in several studies tracking the professional development of specific groups of people, see Chan (1995), Biemann & Datta (2014) or Gubler, Biemann, Tschopp, & Grote (2015), and to analyze life course data, for example, Widmer & Ritschard (2009) or Bastin (2015).

Limitations of optimal matching have been pointed out by some critics. A notable flaw raised by Wu (2000) lies in the definition of the cost matrix — a core hyperparameter of the method — and the strong assumption about its symmetry as initially, the method considered substitution costs that were provided by an expert in the context. However, data-based aproaches for the calculation of the cost matrix have been proposed since then, see Studer & Ritschard (2016).

Once a distance matrix has been obtained via optimal matching or any other technique that is applicable to categorical sequences, a distance-based method can

be used depending on the specific interest of the research. Usually, clusters are obtained in order to identify common trajectories through visual inspection, Abbott (1983) highlighted the possibility to use the distance matrix for multidimensional scaling, Gabadinho & Ritschard (2013) proposed a way to identify typical patterns based on the coverage neighborhood of the sequences applied to childbirth histories, Massoni, Olteanu, & Rousset (2009) combined optimal matching and self-organizing maps in the study of career path and employability. In addition, Studer, Ritschard, Gabadinho, & Müller (2011) propose a methodology to analyze how covariates can explain the discrepancy between sequences based on their dissimilarities. However, to our knowledge, there has been no attempt at using categorical sequences for the prediction of other variables.

In this work we are interested in studying the effect of different hyperparameters when using optimal matching to obtain pairwise distances of a group of sequences and subsequently employ them in distance-based methods. Particularly, we consider the effects on both clustering (unsupervised learning) and variable prediction with k-NN (supervised learning). For this purpose we analyze a new real-world dataset that includes categorical sequences, auxiliary information and other variables of interest: data from the "Women 40+ Healthy Aging Study".

1.3 Women 40+ Healthy Aging Study

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 categorical sequence of interest.

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, as well as habitual drinkers, were excluded. 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 in this work is 239.

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

In order to obtain groups of similar sequences, we apply a hierarchical agglomerative clustering analysis using Ward's method to minimize the dispersion within the clusters (Murtagh & Legendre (2014)), for this purpose we use the method `"ward.D2"` of the `hclust` function in `R`.

On the other hand, for prediction we use the function `k.nearest.neighbors` from the `R` package `FastKNN`. 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_{n+1} 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). That is, for a given distance function $d(\cdot, \cdot)$ we predict y_{n+1} as:

$$\hat{y}_{n+1} = \frac{1}{k} \sum_{j=1}^{k} y_{(j)} \tag{1.1}$$

where $j = (1), \ldots, (k)$ index the nearest k neighbors of x_{n+1}:

$$d(x_{(1)}, x_{n+1}) < \cdots < d(x_{(k)}, x_{n+1}) < d(x_{(k+1)}, x_{n+1}) < \cdots < d(x_{(n)}, x_{n+1}) \tag{1.2}$$

There are different choices for the distance function $d(\cdot, \cdot)$. For instance, the Euclidean or Mahalanobis distances are common choices. In our case, in order to analyze categorical sequences, we apply the more general OM algorithm to obtain a disimiliraty matrix.

1.4 The OM algorithm

As mentioned above, optimal matching is a technique widely applied in social sciences for the comparison of categorical sequences. OM uses the Needleman-Wunsch algorithm to identify similarities between biological sequences that are usually represented as strings of characters.

The goal of OM is to find the best possible alignment between two sequences by considering the differences and equivalences between their elements and minimizing the total cost associated. The cost of changing between states of the sequences we are interested in aligning, can be defined in several ways including data-based methods or values supplied by experts in the particular field.

Consider a set of n categorical states $S = \{s_1, \ldots, s_n\}$, we define $X = (x_1, \ldots, x_t)$, a sequence of length $t < \infty$, 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 disimilarity between the sequences X and Y, we define an empty array F of size $(t_X + 1) \times (t_Y + 1)$. Algorithm 1 below shows the initialization and recursion to fill the array F.

Algorithm 1 Optimal matching.

$F(1,1) \leftarrow 0$
for $j \leftarrow 2, t_Y + 1$ **do**
    $F(1,j) \leftarrow F(1,j-1) + d$
end for
for $i \leftarrow 2, t_X + 1$ **do**
    $F(i,1) \leftarrow F(i-1,1) + d$
end for
for $i \leftarrow 2, t_X + 1$ **do**
    for $j \leftarrow 2, t_Y + 1$ **do**
        $F(i,j) \leftarrow \min\{F(i-1,j)+d, F(i,j-1)+d, F(i-1,j-1)+K(y_{i-1}, x_{j-1})\}$
    end for
end for

The value d is the cost of inserting a gap in one of the sequences, also known 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 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 for a set of sequences.

1.4.1 Example

Suppose that S is the alphabet and let $X = \{S, E, N, D\}$ and $Y = \{A, N, D\}$ be two sequences in $\mathbf{S}$.

Further let $d = 2$, 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				

Then, to fill the second row of F we proceed as follows:

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

What yields:

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

Finally, after completing the recursion for the remaining rows, 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.

The `R` packages `TraMineR` (Gabadinho et al. (2011)) and `TraMineRextras` provide several functions to define, analyze and visualize sequential data. In particular, `TraMineR` implements the OM algorithm and offers several methods for computing the cost matrix K and the normalization of the dissimilarity matrix.

1.5 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` to obtain the cost matrix.

1.5.1 Transition rates (`TRATE`):

The substitution cost between states s_i and s_j, $1 \leq i, j \leq n$ is based on the observed frequencies of the transitions between the states and is calculated as:

$$K(s_i, s_j) = c - P(s_i|s_j) - P(s_j|s_i), \tag{1.3}$$

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

The implementation of this method uses a default value of $c = 2$ which results in substitution costs that are close to 2. Additionally, as pointed out by Studer & Ritschard (2016), the calculation of substitution costs can lead to violations of the triangle inequality, meaning that we obtain a dissimilarity instead of a distance measure.

1.5.2 Chi-squared distance (`FUTURE`):

The χ^2-distance is a weighted sum of the squared differences of distribution vector frequencies. The weight is given by the inverse of the proportion of the total time spent in the state, meaning that the differences on rare states have higher weights.

$$K(s_i, s_j) = d_{\chi^2}(\mathbf{P_i}, \mathbf{P_j}) \tag{1.4}$$

$$= \left[\sum_{l=1}^{n} \alpha_l^{-1} \left(P(s_l|s_i) - P(s_l|s_i) \right)^2 \right]^{1/2} \tag{1.5}$$

where $\mathbf{P}_{\cdot} = (P(s_1|s_{\cdot}), \ldots, P(s_n|s_{\cdot}))'$, $\alpha_l = \sum_{h=1}^{n} P(s_l|s_n)$ and $i \neq j$.

It has been shown via simulation, that this distance is particularly sensitive to the time spent in each state but not so much to the order of the states in a sequence (see Studer & Ritschard (2016)).

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

$$K(s_i, s_j) = d_i + d_j, \tag{1.6}$$

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

$$g_i = \frac{1}{f_i}, \qquad \text{for method 'INDEL', } \tag{1.7}$$

$$g_i = \log\left(\frac{2}{1 + f_i}\right), \qquad \text{for method 'INDELSLOG' } \tag{1.8}$$

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 change of state is always lower or equal than deleting and inserting an element (or vice versa). The reason behind is that higher *indel* costs, compared to the substitution costs, produce dissimilarities that are greatly affected by time shifts.
- The Needleman-Wunsch algorithm with constant costs for mismatch is known as Levenshtein distance (Levenshtein (1966)), a string metric widely used in computer science.
- In general, the resulting measure of the algorithm is a dissimilarity. However, if the cost matrix fulfills the triangle inequality, we obtain a distance measure

(Yujian & Bo (2007)).

1.6 Normalization

By design, OM can deal with sequences of different lengths via insertions. However, in cases when the lengths of the sequences differ greatly, 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_X and t_Y, respectively. Let $d(X, Y)$ be the dissimilarity between the sequences X and Y, t_{max} the length of the longest sequence in $\mathbf{S}$ and d_{max} the maximum dissimilarity between any pair of sequences in $\mathbf{S}$.

`TraMineR` offers the following options to normalize the dissimilarities between sequences:

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

Section 2

Data from the 40+ Healthy Aging Study

2.1 About the data

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

2.2 OM Application

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, we name this the *base scenario*. The cost matrix obtained this way is shown in table 2.1.

As expected, the elements in the diagonal are equal to 0, meaning there is no cost

Table 2.1: Cost matrix obtained from transition probabilities (method 'TRATE') of the relationship data of women over 40 years old.

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

associated to staying in the same state. By default, the constant c in 1.3 is set to 2. This, and the fact that the duration of the states is often longer than 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 observed in the data (e.g. from single without children to single with children and vice versa). Finally, note that we consider missing value (`NA`) 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, or when the length of the sequences differ by a large number.

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

Having obtained the dissimilarities matrix, we can obtain the clusters. Our aim is to explore the data and the differences captured by the dissimilarities matrix. In particular, we set the number of clusters to four, given that this is the maximum number of clusters that produces groups with meaningful differences in the distribution of states. Figure 2.2 shows the distribution of the states for each of the four identified clusters.

In figure 2.3 we show the transverse entropy by cluster, i.e. the cross-sectional

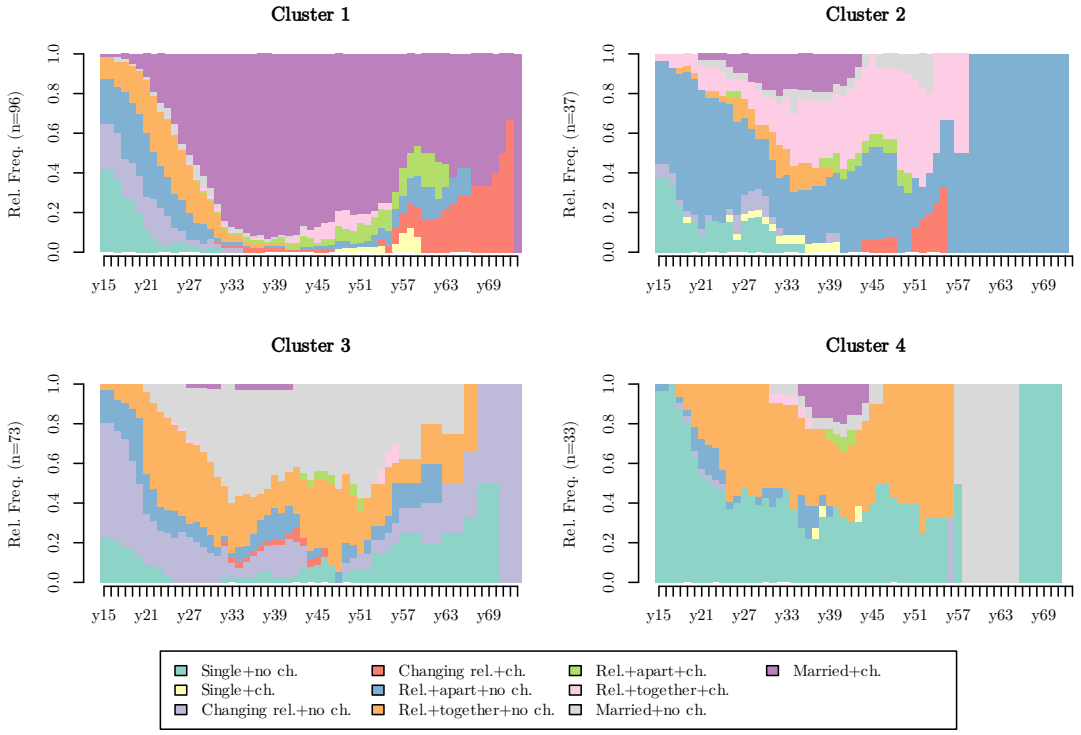

Figure 2.2: Cross-sectional distribution of relationship states by cluster for the base scenario.

entropy of the states distributions is calculated at each time point as follows:

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

These visualizations allow us to identify common and contrasting features of the clusters that can be useful to describe them concisely. In table 2.4 we present a summary of the clusters with our descriptive interpretation of the clusters. It is important to remember that this descriptions are subjective and not exhaustive.

On the other hand, in figure 2.2 we can also appreciate that the conformation of some clusters seems to be highly affected by the length of the sequence and it is reasonable to assume that the normalization method is not achieving the ideal result.

Now, we are interested in exploring how the relationships history of the women relate to personality traits. As a first exploratory step, figure 2.5 shows the distribution of the score for each trait 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 identify differences. For that reason, we also explore with a lower number of clusters.

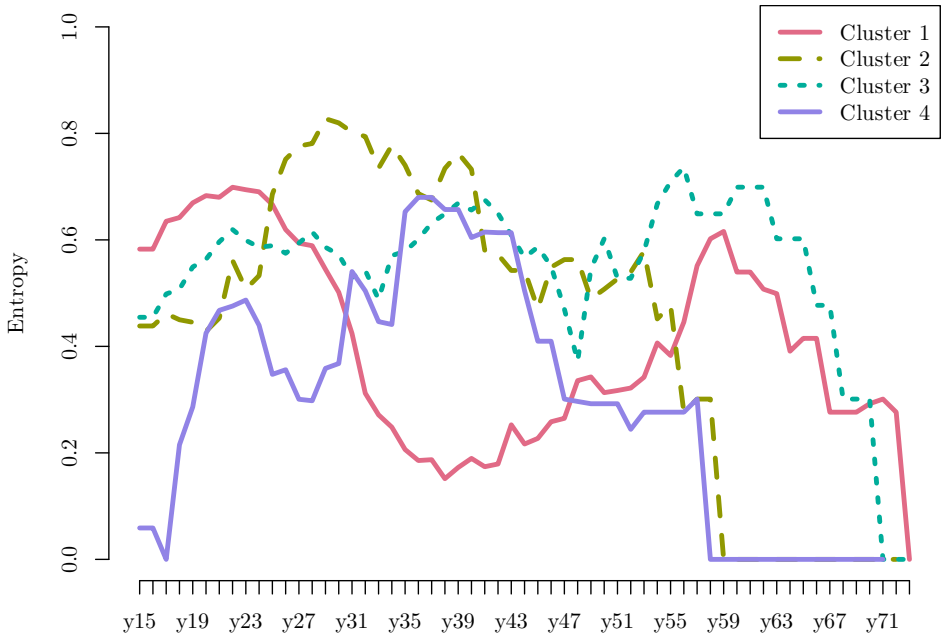

Figure 2.3: Transversal entropy by cluster for the base scenario.

Table 2.4: Summary of the four clusters obtained in the base scenario.

Cluster	n	Share	Description
1	96	40.2%	Married young and had children.
2	37	15.5%	Often in relationships but not married.
3	73	30.5%	Older, mostly married or in long relationship without children.
4	33	13.8%	Younger, single or in a relationship without children.

By setting the number of groups to only two, 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 (figure 2.6): the majority of women in cluster 1 have children, while we mostly find women without children in cluster 2. In addition, the transversal entropy of the sequences for the two clusters is displayed in figure 2.7, where it is shown 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. Figure 2.5 shows the distribution of the personality traits for the two clusters.

There seems to be differences in the distributions of some personality scores: the

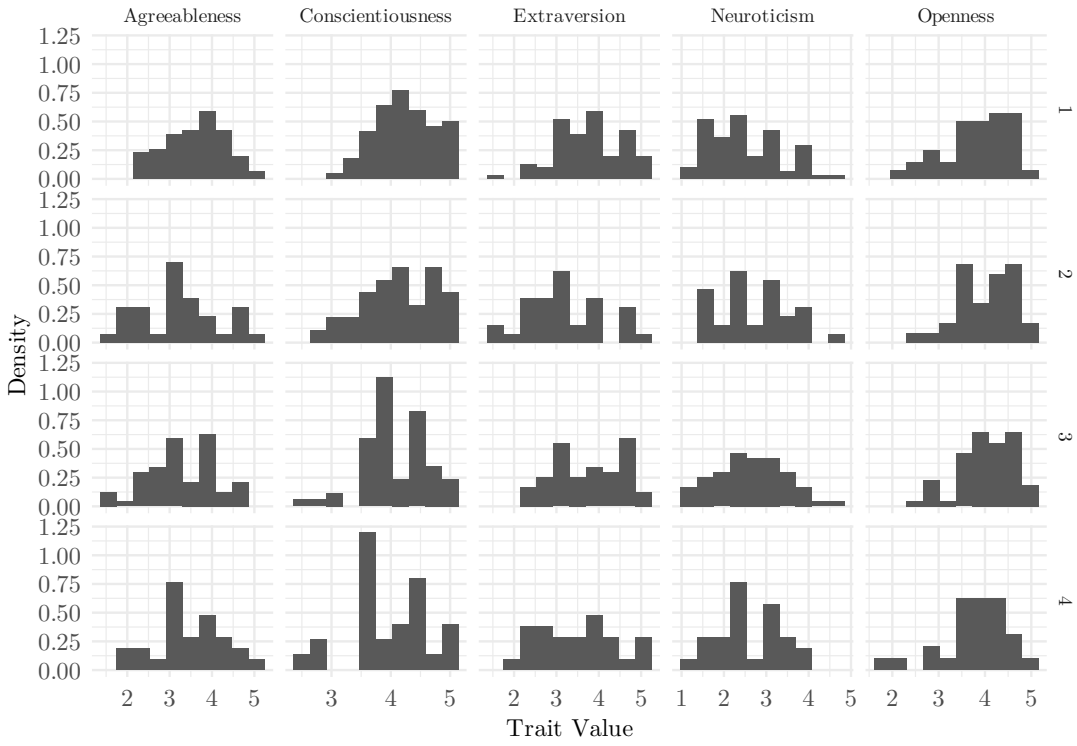

Figure 2.5: Histogram of the five personality scores by cluster in the base scenario.

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, having a dissimilarity matrix provides a numerical expression of the categorical sequences that is useful for other purposes. In particular, in the next section, we explore the predictive capability of this dataset using a non-parametric prediction method.

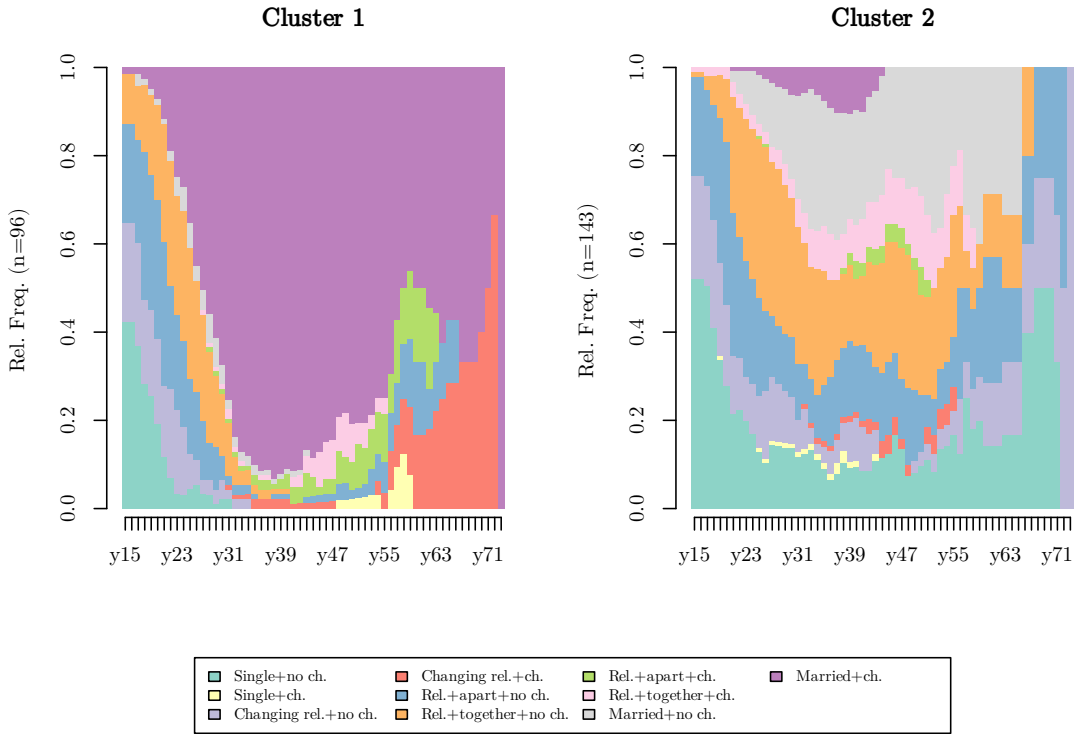

Figure 2.6: Cross-sectional distribution of states for two clusters in the base scenario.

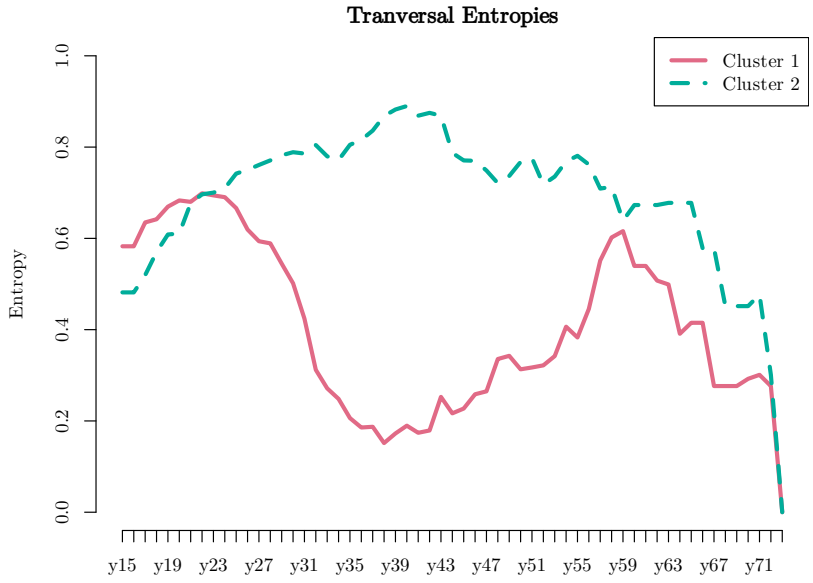

Figure 2.7: Transversal entropy for two clusters in the base scenario.

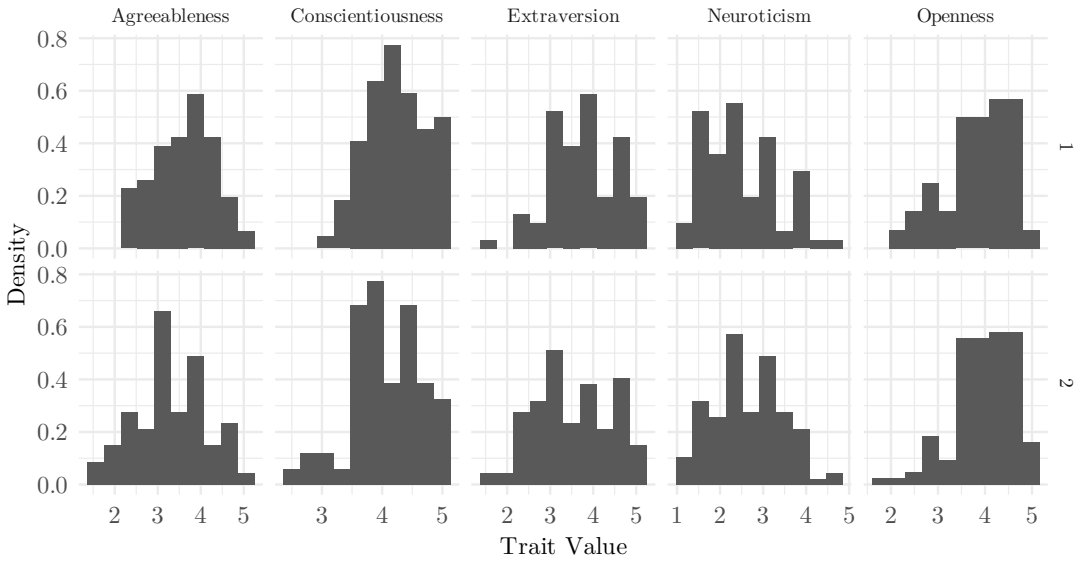

Figure 2.8: Histogram of the five personality scores for two clusters in the base scenario.

Section 3

Personality Scores Prediction with k-Nearest Neighbors

3.1 k-Nearest Neighbors

Note that in this setup, $d(\cdot,\cdot)$ is one of the multiple normalized dissimilarities $d(\cdot,\cdot) = d^{'}(\cdot,\cdot|K)$ and it is dependent or parametrized by the cost matrix K.

k is also 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 capture better the local structure.

To compare the performance of different values of k and other hyperparameters, we use the mean squared error (MSE). For a testing set of m labeled data points, the MSE is given by:

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

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

3.2 Personality scores prediction in base scenario

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

For each trait we predict the personality score values Y and compare them with the observed values using the MSE. Table 3.1 summarizes the results of the optimal prediction and Figure 3.2 shows the MSE for the different values of k, i.e. for $k = 1, \ldots, 80$. As a reference, a red line for every personality trait is added to indicate the MSE of the trivial prediction, i.e. the mean of all the sample points in the train set. Recall that we are using the cost matrix K shown in Table 2.1 and normalized dissimilarity defined by the `maxlength` method as described in Subsection 2.2.

Table 3.1: Summary of k-nn prediction of personality scores in the base scenario.

Trait	min(MSE)	k	Trivial MSE
Agreeableness	0.59	48	0.62
Conscientiousness	0.37	15	0.40
Extraversion	0.71	22	0.76
Neuroticism	0.64	7	0.71
Openness	0.54	12	0.54

Overall, it seems that using the sequential data for prediction results in little improvement compared to the trivial prediction. For neuroticism, the MSE decrases rapidly, reaching the lowest value at $k = NA$ and then increases again.

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 for $k = 15$ and after $k = 30$ the MSE curve stays flat.

For agreeableness, the MSE increases again after the optimal k. However, note that this minimum is not considerably lower than the trivial prediction. Similarly, for extraversion, the MSE takes a minimum value with $k = 22$, 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, namely the cost matrix and choice of normalized dissimilarity.

3.3 Additional scenarios considered for prediction

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

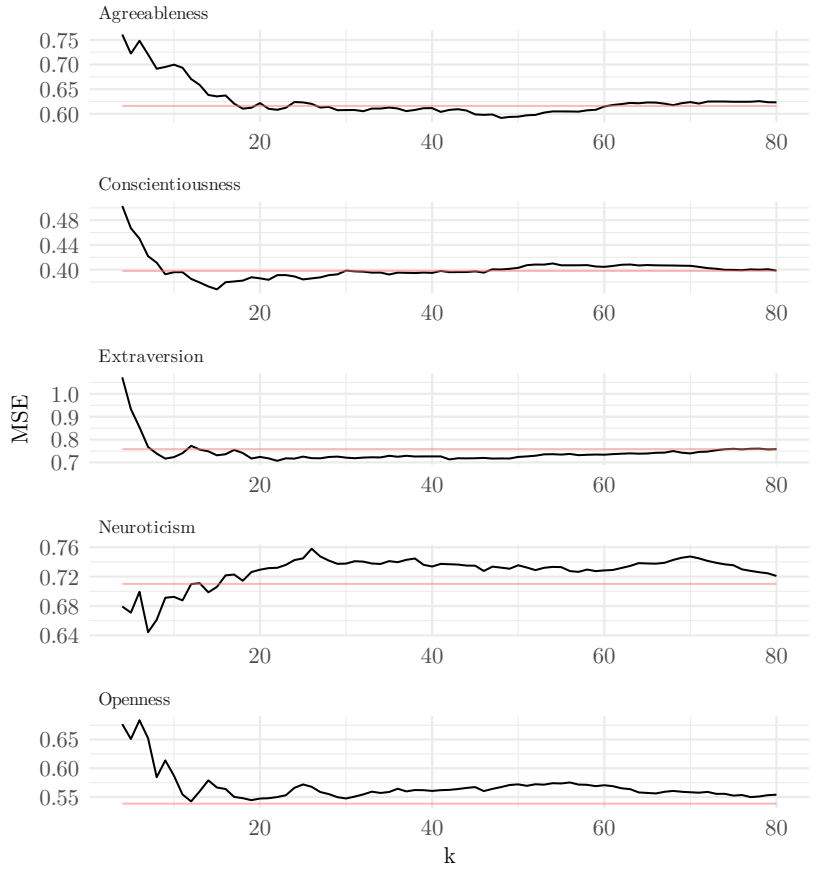

Figure 3.2: Prediction MSE by personality trait in the base scenario with trivial prediction reference (red line).

- Take the constant c in 1.3 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 Figure 2.2).
- 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.

Table 3.3: Summary of additional scenarios considered for obtaining dissimilarity matrix.

	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 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.

$$p = (1 - (MSE_k / MSE_{trivial})) * 100 \tag{3.2}$$

Figure 3.4 shows the best relative improvement achieved for each personality trait and under all the scenarios in 3.3 and the corresponding value of k at which the best performance was obtained. We can observe that there is not a single scenario that

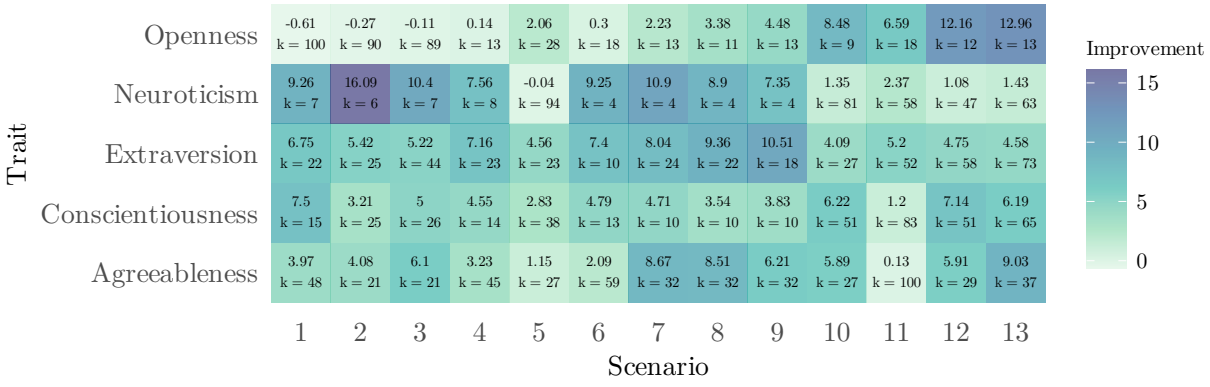

Figure 3.4: Improvement of MSE in the prediction of personality scores relative to the trivial prediction and respective value of k for each scenario.

produces the best prediction improvement for every trait. Although, the method `FUTURE` seems to produce better results for all of the traits except neuroticism.

Figure 3.5 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. We can observe that the best prediction

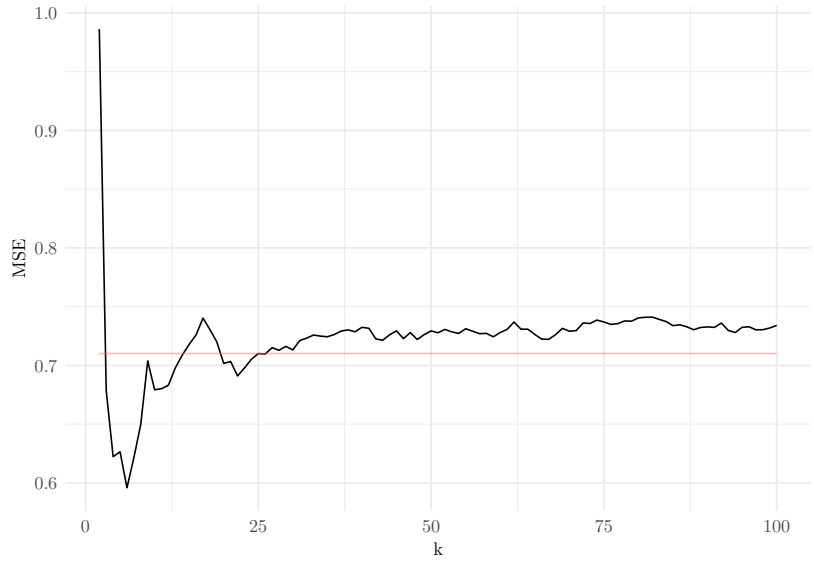

Figure 3.5: MSE of neuroticism prediction with trivial prediction MSE as reference (red line) in scenario 2.

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 3.4 we observe the minimum MSE is found with a similar number of neighbors for other scenarios, for instance, in scenario 3 (see Figure 3.6). For openness, the best predictions are obtained with scenarios 12 and 13. In both scenarios the cost matrix is calculated with χ^2 distance of the states frequencies (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. Figure 3.7 shows the MSE for openness 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 Figure 3.8, the MSE for this scenario is shown. 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

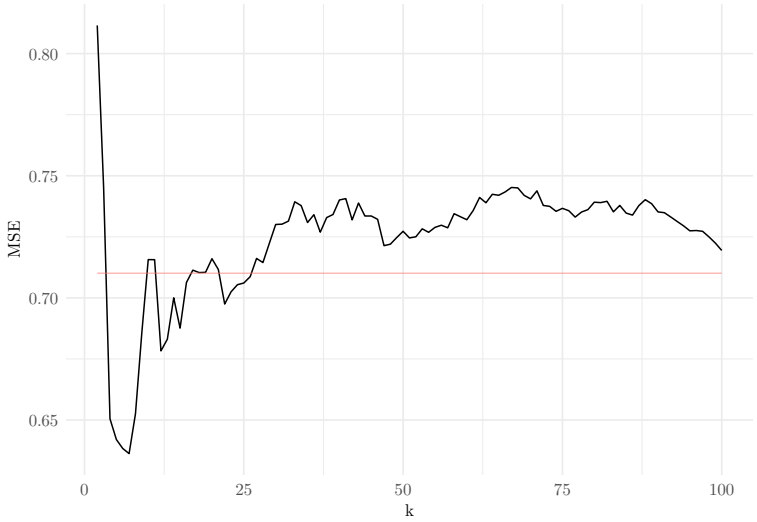

Figure 3.6: MSE of neuroticism prediction with trivial prediction MSE as reference (red line) in scenario 3.

at least 10%. Furthermore, the best prediction is obtained for $k = 15$ in the base scenario. Scenario 12 produces a similar result in terms of improvement, but with $k = 51$ which is an undesirable high number of neighbors for prediction with kNN. Figure 3.9 shows in detail the MSE for conscientiousness in the base scenario. Likewise, for agreeableness, all of the scenarios showed improvements relative to the trivial prediction that are below 10% and the minimum MSE in every case is obtained for rather 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 Figure 3.10 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. Finally, we perform clustering again with the distance matrix obtained in Scenario 13. Table 3.11 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 (excluding the diagonal elements and the missing value cost) is larger than those observe in 2.1. Figure 3.12 shows the distribution of states by cluster for this scenario. Even tough the cost matrix presented large variations compared to the base scenario and the predictions improved, we obtain clusters that exhibit similar main characteristics: In cluster 1, we find women with different relationship situations but without children. Similarly, in cluster 2 we find women with different relationship trajectories, but mostly married, that eventually had children. The clusters seem to be better defined in this case but that can be also due to the age restriction imposed, which implies that some sequences without enough data in the specified age range

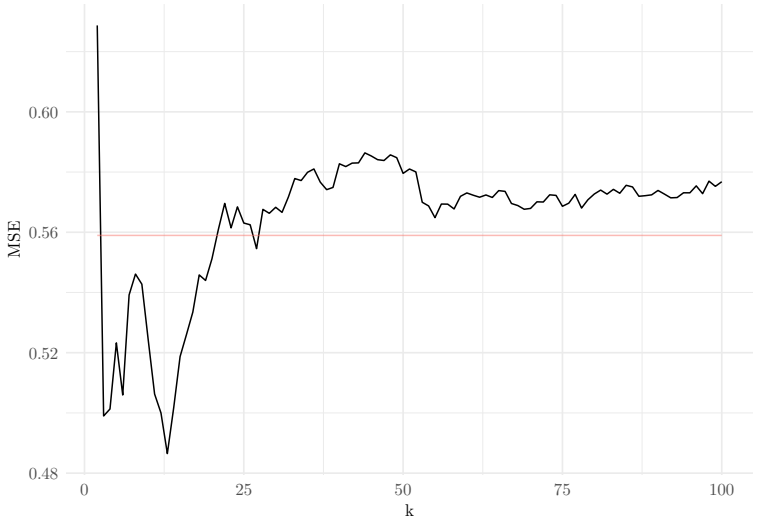

Figure 3.7: MSE of openness prediction with trivial prediction MSE as reference (red line) in scenario 13.

were excluded.

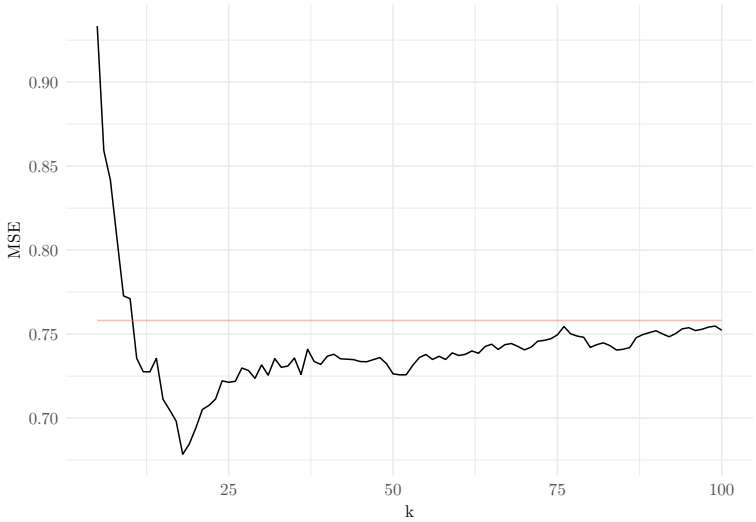

Figure 3.8: MSE of extraversion prediction with trivial prediction MSE as reference (red line) in scenario 9.

Figure 3.9: MSE of conscientiousness prediction with trivial prediction MSE as reference (red line) in scenario 2.

Figure 3.10: MSE of agreeableness prediction with trivial prediction MSE as reference (red line) in scenario 13.

Table 3.11: Cost matrix for scenario 13: sequences restricted to ages from 20 to 55, cost from/to missing value of 0.5, method for calculation of cost matrix 'FUTURE' and dissimilarity normalization 'gmean'.

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

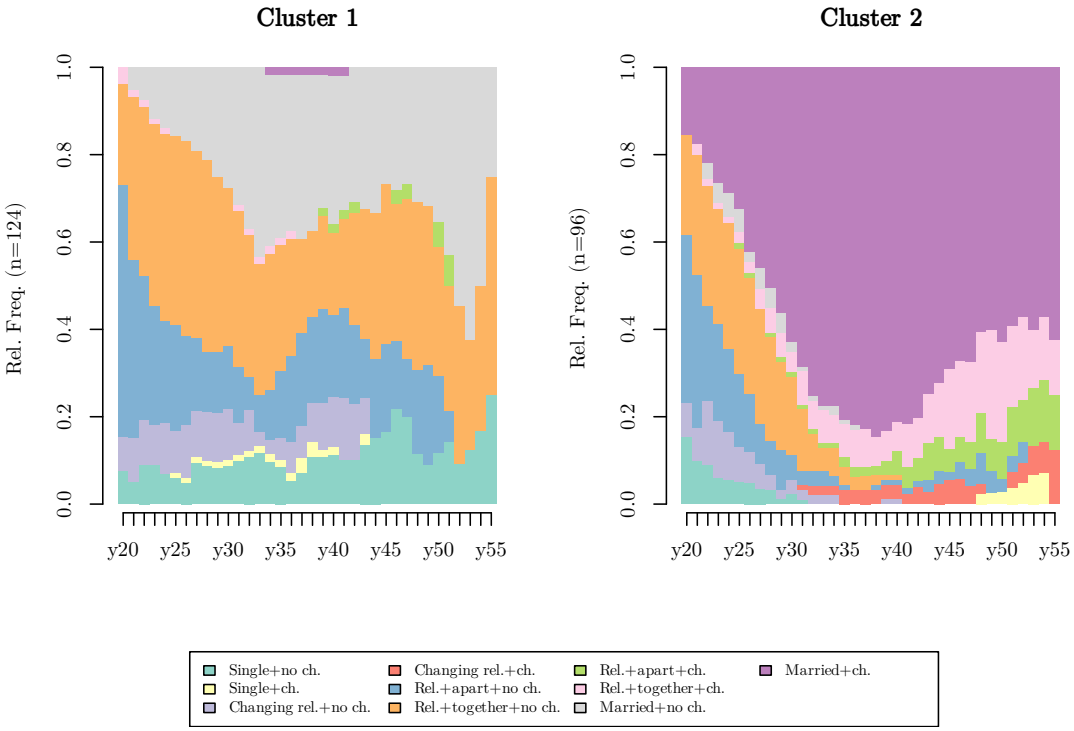

Figure 3.12: Cross-sectional distribution of states for two clusters in Scenario 13.

Conclusion

The use of optimal matching to describe sequential data in the context of social studies has been extensively documented. We explored the potential to use this technique for prediction. Despite the lack of a measure to directly compare cost or distance matrices produced by optimal matching, we decided to use MSE to evaluate different scenarios as our main goal was to obtain good performance with the predictions.

Based on the analysis of the distance matrix obtained through optimal matching, we have determined that there is not a single specific combination of cost matrix generation method, normalization method, and treatment of missing values that consistently yields the best prediction for personality scores. However, when it comes to most personality traits, it is preferable to use the cost matrix calculated with the χ^2 distance of the states frequencies. It is worth noting that critics of this technique have pointed out that the results heavily rely on how the cost matrix is defined and we have confirmed this with our results.

We also noted that the difference in sequence lengths affected the performance of the prediction as this produced a large amount of missing values, showing the limitation of any normalization method used.

Furthermore, we also observed that in the two scenarios that we considered for clustering, the resulting states distributions when dividing the individuals into two clusters, show that the main differentiating factor between them is the presence or absence of children (see Figure 2.6 and Figure 3.12).

As for future work, we consider that it would be convenient to consider another method for obtaining the cost matrix. For instance, in the specific context of the relationship history, it makes sense to consider an asymmetrical cost matrix as it is likely that the change from one state to another is more frequent that the converse.

References

Abbott, A. (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. http://doi.org/10.1080/01615440.1983.10594107

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`

Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods & Research*, *29*(1), 3–33. http://doi.org/10.1177/0049124100029001001

Bastin, S. (2015). *Partnerschaftsverläufe alleinerziehender mütter: Eine quantitative untersuchung auf basis des beziehungs- und familienpanels*. Springer Fachmedien Wiesbaden. Retrieved from `https://books.google.ch/books?id=ttMjCgAAQBAJ`

Bergroth, L., Hakonen, H., & Raita, T. (2000). A survey of longest common subsequence algorithms. In *Proceedings seventh international symposium on string processing and information retrieval. SPIRE 2000* (pp. 39–48). http://doi.org/10.1109/SPIRE.2000.878178

Biemann, T., & Datta, D. K. (2014). Analyzing sequence data: Optimal matching in management research. *Organizational Research Methods*, *17*(1), 51–76. http://doi.org/10.1177/1094428113499408

Chan, T. W. (1995). Optimal matching analysis: A methodological note on studying career mobility. *Work and Occupations*, *22*(4), 467–490. http://doi.org/10.1177/0730888495022004005

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, *7*(3), 171–176.

Gabadinho, A., & Ritschard, G. (2013). Searching for typical life trajectories applied to childbirth histories. In (pp. 287–312).

Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, *40*(4), 1–37. http://doi.org/10.18637/jss.v040.i04

Gubler, M., Biemann, T., Tschopp, C., & Grote, G. (2015). How career anchors differentiate managerial career trajectories: A sequence analysis perspective. *Journal of Career Development*, *42*(5), 412–430. http://doi.org/10.1177/0894845315572891

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell System Technical Journal*, *29*(2), 147–160. http://doi.org/10.1002/j.1538-7305.1950.tb00463.x

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction* (2nd ed.). Springer. Retrieved from `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`

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

Massoni, S., Olteanu, M., & Rousset, P. (2009). Career-path analysis using optimal matching and self-organizing maps. In. http://doi.org/10.1007/978-3-642-02397-2_18

Murtagh, F., & Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, *31*(3), 274–295. http://doi.org/10.1007/s00357-014-9161-z

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`

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from `https://www.R-project.org/`

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, *179*, 481–511. http://doi.org/10.1111/rssa.12125

Studer, M., Ritschard, G., Gabadinho, A., & Müller, N. S. (2011). Discrepancy analysis of state sequences. *Sociological Methods & Research*, *40*(3), 471–510. http://doi.org/10.1177/0049124111415372

Widmer, E. D., & Ritschard, G. (2009). The de-standardization of the life course: Are men and women equal? *Advances in Life Course Research*, *14*(1), 28–39.

http://doi.org/https://doi.org/10.1016/j.alcr.2009.04.001

Winkler, W. (1990). String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*.

Wu, L. L. (2000). Some comments on "sequence analysis and optimal matching methods in sociology: Review and prospect." *Sociological Methods & Research*, *29*(1), 41–64. http://doi.org/10.1177/0049124100029001003

Yujian, L., & Bo, L. (2007). A normalized levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(6), 1091–1095. http://doi.org/10.1109/TPAMI.2007.1078