UNIVERSITÁ COMMERCIALE LUIGI BOCCONI - MILANO

Facoltá di Economia

Dottorato di Ricerca in Statistica
XX Ciclo

New methods for description and prediction in sequence analysis

Coordinatore:
Ch.mo Prof. Pietro Muliere

Tesi di: Gaia Salford
Matricola: P1003680

UNIVERSITÁ COMMERCIALE "LUIGI BOCCONI"
ISTITUTO DI METODI QUANTITATIVI

The thesis "**New methods for description and prediction in sequence analysis**" by **Gaia Salford** is recommended for acceptance by the members of the delegated committee, as stated by the enclosed reports, in partial fulfillment of the requirements for the degree of **Doctor of Philosophy**.

Dated: January 2008

Research Supervisor: **Ch.mo Prof. Francesco Billari**

External Examiners: **Marco Bonetti**
Cees Elzinga

UNIVERSITÁ COMMERCIALE "LUIGI BOCCONI"

Date: **January 2008**

Author: **Gaia Salford**

Title: **New methods for description and prediction in sequence analysis**

Department: **Istituto di Metodi Quantitativi**

Signature of Author

Table of Contents

Table of Contents vii

List of Tables ix

List of Figures xiv

1 Sequence analysis 1
- 1.1 Introduction . . . 1
- 1.2 The motivating data . . . 3
- 1.3 Metric representation of categorical time series . . . 5
  - 1.3.1 Elzinga's metrics . . . 8
- 1.4 Cluster and MDS: a graphical tool . . . 11
- 1.5 Explaining and predicting life courses: the methods . . . 22
- 1.6 Appendix . . . 26

2 Predicting teenage and long run non-marital childbearing using sequence based methods 29
- 2.1 Introduction . . . 29
- 2.2 Data and methods . . . 32
  - 2.2.1 Data . . . 32
  - 2.2.2 Explanatory variables . . . 33
  - 2.2.3 Methods . . . 34
- 2.3 Results and discussion . . . 37
- 2.4 Concluding remarks . . . 45
- 2.5 Appendix . . . 47

3 Analysis of dispersion (ANODI) with permutation test 51
- 3.1 Introduction . . . 51
- 3.2 Analysis of dispersion and permutation test . . . 53
- 3.3 Simulation study . . . 55

3.4 Model selection and permutation test 62
3.5 ANODI of the FFS data 64
3.6 Conclusions 76

4 A parametric approach to sequence analysis 79
4.1 Introduction 79
4.2 The model 82
4.3 Analysis of FFS Data 84
4.4 Conclusion and remarks 97
4.5 Appendix 100

Conclusions 105

Bibliography 109

List of Tables

1.1 Description for the obtained cluster solution 18

1.2 Logit model coefficients for the six-clusters solution (Boldface indicates p-values lower than 0.05). 23

2.1 Sample size for each category; 34

2.2 Fractional logit model-coefficients for Sweden. 38

2.3 Fractional logit model-coefficients for Finland. 38

2.4 Fractional logit model-coefficients for USA. 39

2.5 Fractional logit model-coefficients for Canada. 39

2.6 Fractional logit model-coefficients for Poland. 40

2.7 Fractional logit model-coefficients for Estonia. 40

2.8 Fractional logit model-coefficients for Sweden. 41

2.9 Fractional logit model-coefficients for Italy. 41

2.10 Fractional logit model-coefficients for Spain. 42

2.11 Fractional logit and beta regression model-coefficients for Sweden. 47

2.12 Fractional logit and beta regression model-coefficients for Finland. 47
2.13 Fractional logit and beta regression model-coefficients for USA. 48
2.14 Fractional logit and beta regression model-coefficients for Canada. 48
2.15 Fractional logit and beta regression model-coefficients for Poland. 49
2.16 Fractional logit and beta regression model-coefficients for Estonia. 49
2.17 Fractional logit and beta regression model-coefficients for Sweden. 50
2.18 Fractional logit and beta regression model-coefficients for Italy. 50
2.19 Fractional logit and beta regression model-coefficients for Spain. 50

3.1 Sample size of the covariates for Dutch FFS data. 65
3.2 Summary table for ANODI-permutation test on Dutch FFS data. 65
3.3 Selection procedure based on partial effects. Variable are: 72

4.1 80 per cent confidence interval coverage probabilities based on 100 simulated samples. 85
4.2 The Netherlands covariates' description. 86
4.3 Number of total transitions in FFS Dutch data. 87
4.4 Number of total transitions in FFS Dutch data. 88
4.5 Number of transitions (per cent) in FFS Dutch data. . . 88
4.6 Number of transitions (percent) in FFS Dutch data. . . . 88
4.7 Visited states' description. 89
4.8 Summary of sequential likelihood-ratio test to get a final parsimonious model (see text for the definition of the covariates). 90
4.9 Parameters estimates for the selected model. 91
4.10 Comparison between the mean of the observed times in the first state, the mean of times simulated with our model without censoring after 144 months, and with censoring, given each covariates' combination. 93
4.11 Comparison between the mean of the observed times in the first state and the mean of times simulated with our model, the expected value estimated, with and without censoring, given each covariate. 93
4.12 Comparison between the mean of observed and censored simulated permanence times for the second, the third, and fourth visited state, given each covariates' combination. . 94

4.13 Comparison between the mean of observed and censored simulated permanence times for the second, the third, and fourth visited state, given each covariate. 94
4.14 Summary table for ANODI-permutation test on Dutch FFS data and on a simulated sample under the estimated model; 97
4.15 Comparison between the observed and simulated transitions for women religious, with education level 2, parents not divorced or separated, and belonging to '53-'57 cohort. 101
4.16 Comparison between the observed and simulated transitions for women religious, with education level 3, parents not divorced or separated, and belonging to '53-'57 cohort. 101
4.17 Comparison between the observed and simulated transitions for women religious, with education level 2, parents not divorced or separated, and belonging to '58-'62 cohort. 101
4.18 Comparison between the observed and simulated transitions for women religious, with education level 3, parents not divorced or separated, and belonging to '58-'62 cohort. 102
4.19 Mean of the observed and censored simulated permanence times for two classes of ages at entry into the first, the second and third visited states. 102
4.20 Number of observed and simulated transitions for women that have transitions before *51*-th months from start of observation period (22 years and 3 months old). 105

4.21 Number of observed and simulated transitions for women that have transitions after *51*-th months from start of observation period (22 years and 3 months old). 105
4.22 Number of observed and simulated transitions in which the time spent in the origin state was less than 40 months. 105
4.23 Observed and simulated transitions in which the time spent in the origin state was more than 40 months. . . . 106

List of Figures

1.1 An examples of possible observed life sequence, S/40 U/30 M/74. Time (months) is on the horizontal axis. Refer to the text for the meaning of the states. 4

1.2 Representation of individual sequences of states using first MDS component ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC). . . 15

1.3 Projection of the three-dimensional multidimensional scaling solution on the xy, xz and yz planes. Points are colored according to the cluster membership: black = cluster 1, red = cluster 2, green = cluster 3, blue = cluster 4, sky-blue = cluster 5, pink = cluster 6. 16

1.4 Representation of individual sequences of states, separated in clusters, using first MDS component ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC). 17

1.5 Projection of multidimensional scaling solution. Points are colored according to the number of states visited by individuals: black = 1, red = 2, green = 3, blue = 4, sky-blue = 5, pink = 6, yellow = 7, grey = 8, violet = 9. 19

1.6 Projection of multidimensional scaling solution. Points are colored according to the quartiles of the distribution of the time spent by individuals in state "single" (black = 1^{st} quartile, red = 2^{nd} quartile, green = 3^{rd} quartile, blue = 4^{th} quartile). 19

1.7 Projection of multidimensional scaling solution. Points are colored according to the quartiles of the distribution of the time spent by individuals in state "marriage" (black = 1^{st} quartile, red = 2^{nd} quartile, green = 3^{rd} quartile, blue = 4^{th} quartile). 20

1.8 Projection of multidimensional scaling solution. Points are colored according to the quartiles of the distribution of the time spent by individuals in state "cohabitation" (black = 1^{st} quartile, red = 2^{nd} quartile, green = 3^{rd} quartile, blue = 4^{th} quartile). 20

1.9 Projection of multidimensional scaling solution. Points are colored according to the quartiles of the distribution of the time spent by individuals in state "single with children" (black = 1^{st} quartile, red = 2^{nd} quartile, green = 3^{rd} quartile, blue = 4^{th} quartile). 21

1.10 Projection of multidimensional scaling solution. Points are colored according to the quartiles of the distribution of the time spent by individuals in state "marriage with children" (black = 1^{st} quartile, red = 2^{nd} quartile, green = 3^{rd} quartile, blue = 4^{th} quartile). 21

1.11 Projection of multidimensional scaling solution. Points are colored according to the quartiles of the distribution of the time spent by individuals in state "cohabitation with children" (black = 1^{st} quartile, red = 2^{nd} quartile, green = 3^{rd} quartile, blue = 4^{th} quartile). 22

1.12 Representation of individual sequences of states, separated in clusters, using MDS ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC). 24

1.13 Representation of individual sequences of states, separated according to estimated clusters, using MDS ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC). 25

1.14 Representation of individual sequences of states, separated in clusters, using first three MDS component ordering (NCS with duration as weights). Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC). 27

1.15 Representation of individual sequences of states, separated in clusters, using first three MDS component ordering (OM with with symmetric cost matrix). Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC). 28

2.1 Beta densities for different combinations of (α, β). 35

2.2 Histograms of the distance from SC/145 template, computed with different metrics for the Netherlands. 43

2.3 Representation of individual sequences ordered according to the distance from the template SC/145. Time in months is on the vertical axis and the number of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC). 44

3.1 Histograms of the p-values obtained with ANODI test with "structured" and "unstructured" sequences, for 1000 trials based on 2000 permutated samples. 60

3.2 Histograms of the p-values obtained in 1000 trials with ANODI test based on 2000 permutated samples from "structured" and "unstructured" sequences when the number of groups of the covariate is 3, 4, and 5. 62

3.3 Histograms of the p-values obtained in 1000 trials with ANODI test based on 2000 permutated samples from "structured" and "unstructured" sequences for different probabilities of belonging to the two groups of the covariate: P=(0.5, 0.5), P=(0.75, 0.25) and P=(0.95, 0.05). 63

3.4 Permutation distribution of the statistic $T = B_D/W_D$ based on the categories of "Educ"-covariate of Dutch FFS data. The dashed vertical line marks the observed value of the statistic, T_0. 66

3.5 Permutation distribution of the statistic $T = B_D/W_D$ based on categories of "Religion"-covariate of Dutch FFS data. The dashed vertical line marks the observed value of the statistic, T_0. 66

3.6 Permutation distribution of the statistic $T = B_D/W_D$ based on categories of "Divorce"-covariate of Dutch FFS data. The dashed vertical line marks the observed value of the statistic, T_0. 67

3.7 Permutation distribution of the statistic $T = B_D/W_D$ based on categories of "Cohort"-covariate of Dutch FFS data. The dashed vertical line marks the observed value of the statistic, T_0. 67

3.8 Representation of individual sequences of states, separated in education' s level groups, using MDS components ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC.) 69
3.9 Representation of individual sequences of states, separated in religious and not religious women, using MDS components ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC). 70
3.10 Representation of individual sequences of states, separated in groups with or without parents divorced, using MDS components ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC). . . 71
3.11 Representation of individual sequences of states, separated for cohort birth, using MDS components ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC). 73

3.12 Representation of individual sequences of states, separated in groups induced by "Educ" and "Religion" (column 1) and by "Educ", "Religion" and "Cohort" (columns 2-3), using MDS components ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC). 74

3.13 Representation of individual sequences of states, separated in groups induced by "Religion" (column 1) and by "Religion" and "Educ" (columns 2-3), using MDS components ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC). 75

3.14 Representation of individual sequences of states, separated in groups induced by "Educ" (column 1) and by "Educ" and "Religion" (columns 2-3), using MDS components ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC). 76

4.1 Three examples of possible observed life sequences (S: living single without children, M: living married without children, U: living in unmarried cohabitation without children, C: with at least one child. Time (months) is on the horizontal axis. 87
4.2 Observed (solid) and simulated (boldface) survival distributions for the time in the first visited state. 95
4.3 Observed (solid) and simulated (boldface) survival distributions for the time in the second visited state 96
4.4 Observed (solid) and simulated (boldface) survival distributions for the time in the third visited state 97
4.5 Observed (solid) and simulated (boldface) survival distributions for the first birth event 98
4.6 Observed (solid) and simulated (boldface) survival distributions for the first union event (cohabitation or marriage) 99
4.7 Observed (solid) and simulated (boldface) survival distributions for the first marriage event 100

Acknowledgements

I would like to thank many people who made this thesis possible.
First of all, my supervisors. Prof. F. C. Billari, for his helpful guidance and advices when taking difficult decisions. He has inspired me the topic of this thesis and I owe him a number of fruitful and crucial advices to the successful completion of the thesis. Prof. M. Bonetti for his constant support and many suggestions during this research. His guidance was fundamental. Thanks to his patience and his enthusiasm for the research I overcame trouble moments. Thank you for your knowledge, patience and generosity with your time. Prof. R. Piccarreta for her precious technical and human support. She has led me patiently in the development of the work with brilliant ideas and many valuable suggestions. I thank you for the interesting and funny discussion on sequence analysis.
I am also very grateful to Prof. C. Elzinga for the time and the attention he devoted to me during my visiting period at VU University of Amsterdam. He provided me with a number of useful and accurate remarks and suggestions.
Thanks also to all the professors and the mates of my Ph.D. course for the help, support and encouragement throughout this "trip".
And finally, I must thank my parents. Thank you for 27 years of minding me. I promise I will get a job one day!
If I have forgotten anyone my apologies, but thanks to everyone who helped me during this PhD.

Chapter 1

Sequence analysis

1.1 Introduction

The life course approach to the study of demographic behavior has attracted increasing interest in the recent literature. One of the techniques that have been proposed to analyse life course data is *sequence analysis*. In sequence analysis each individual's life course is considered as a sequence of states by an holistic point of view (Billari, 2001), i.e. life courses are thought of as being the outcomes of life-planning and it is thus meaningful to consider each sequence as a whole in the input to statistical analyses. The analysis of such data is very complex and proper multivariate statistical techniques are necessary to extract information from such elaborate structure. In the literature, cluster analysis has become the major approach and different distance measures and classification techniques were proposed in order to find and efficiently describe the set of ideal-type sequences observed (see the reviews in Abbott, 1995 and Abbott, Tsay, 2000). Abbott and Forrester (1986) first applied Optimal Matching to calculate distances between trajectories and used them to cluster life courses in different groups; Billari and Piccarreta (2001,2005) proposed a monothetic divisive algorithm, using OM distance, that resulted in a top-down hierarchical clustering. It divided the data into smaller groups one variable at a time, making the clustering process

explicit and easy to interpret. Billari, Fürnkranz and Prskawetz (2006) used a machine learning technique that allowed the detection of the characteristics which distinguish different sets of individuals through a decision tree. Elzinga (2005, 2006) introduced some new metrics and relative similarity measures to compute the distance between sequences. We will present some of these metrics with more details in Section 4. A nonparametric approach for formal hypothesis testing was proposed by Kowalski, De Gruttola, Pagano (2001) in the field of genetics. They presented a test for differences between groups of highly dimensional genetic sequences in comparison of the interpoint distance distributions within groups. The comparison is based on the M-statistic (see Bonetti and Pagano, 2004).

A problem, that is still open concerns the explanation of sequences on the basis of a set of explanatory variables. Until now, the principal contribution is given by McVicar and Anyadike-Danes (2002). They suggested to use clusters obtained with Optimal Matching distance as a dependent variable in a multinomial logit model to study the connection between sequences and some socio-economic variables. In Section 5 we show that the McVicar and Anyadike-Danes' model applied to Dutch FFS data performs quite poorly as far as forecasting is concerned.

The lack of results presently available make explanation and prediction of sequences particularly challenging analysis problems. In this thesis, in particular, our goals are:

- to identify and evaluate which individual and family factors explain the dissimilarity between life course trajectories;
- to model the whole process that generates sequences taking the explanatory structure into account, in order to estimate the effect of the explanatory variables and also to predict life courses for any given combination of the explanatory variables;
- to use the output distance to detect and predict sub-groups as potential targets

for specific public policies.

We propose three methods that we will present briefly in Section 5 and which we will develop in depth in the following chapters. They will be presented as three distinct projects.

The remainder of Chapter 1 is organized as follows. Section 2 provides a description of the data. In Section 3 we describe and evaluate several metrics for the representation of categorical time series. In Section 4 we present a graphical instrument to depict the structure of sequences based on Dutch FFS data. Section 5 briefly describes methodological issues for explanatory and predictive analysis.

1.2 The motivating data

The data for this analysis are from the Family and Fertility Survey (FFS) that was conducted between 1988 and 1999 by the Population Activities Unit (PAU) of the United Nations Economic Commission for Europe (UNECE) in order to have a better understanding of recent trends, current patterns and the possible course of family-related behaviour. The study was carried out in collaboration with 23 UNECE countries plus New Zealand. In this chapter, in Chapter 3 and 4 we will focus only on the Dutch data, while in Chapter 2 we will adopt an international comparative approach and we will analyse the data from Italy, Spain, the Netherlands, Poland, Estonia, Sweden, Finland, USA and Canada.

Dutch FFS was conducted in 1993. It is based on a retrospective sample of 3700 men and 4500 women living in the Netherlands and born in the period 1950-1974. Here we only use the female sample, born from 1953 and 1962. The available dataset contains 1893 women.

In the FFS, retrospective histories of childbearing and family formation of women were

collected on a monthly time scale. In what follows, we focus on ages between 18 and 30 years, leading to a trajectory of 144 consecutive states for each woman in the sample. The state are: living single without children (S), married without children (M), in unmarried cohabitation without children (U), single with at least one child (SC), married with at least one child (MC), in unmarried cohabitation with at least one child (UC). Note that each state might be repeated more than once for each individual, but once an individual visits a "Children state" she can not return in a "without-Children state". Let us represent, for example, the life course of a woman has been single for 40 months, then cohabitated for 30 months and finally got married living with her husband for 74 months via the sequence S/40 U/30 M/74. Figure 1.1 shows this life trajectory.

The questionnaire covered many aspects of the women's life. In particular we focus on

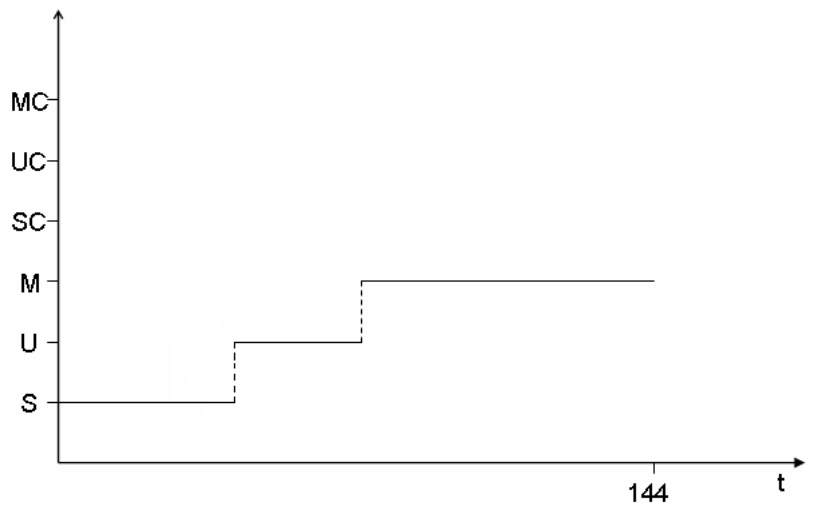

Figure 1.1: An examples of possible observed life sequence, S/40 U/30 M/74. Time (months) is on the horizontal axis. Refer to the text for the meaning of the states.

the level of education, the birth cohort, religiousness of the women and whether their

parents were separated or divorced, at the moment of the interview.

1. The level of education attained of the women (Education), encoded as 1, 2 or 3 if the woman has no education, from 0 to 3 years of education or more than 3 years of education after age of 15. Education data is not available for Canada.

2. Religiousness is inserted as a dummy variable indicating whether the woman is religious (Religion). No categorization is done for different religions or for different levels of participation in religious activities. Data on religiousness are not available for Sweden and Finland.

3. Parental divorce (Divorce). A dummy variable for family background is included in the model and indicates whether parents are separated or divorced.

4. Cohort (Cohort). We used two cohorts: born between 1953 and 1957 or between 1958 and 1962. The first cohort serves as the reference in the model.

Because the informations on the women were collected at the moment of the interview (after the age of 30), Religion and Divorce could not be used to explain or predict the sequences. Education and Cohort have not the same problem because the first variable pertains to a status before the beginning of the sequences and the other does not vary over time. Considering that most of parental divorces take place during adolescence, we believe that to insert Divorce does not cause substantial distortions. The choice of including Religion was more problematic due to the possibility to change the religion belief during the life. Despite that, we believe that Religion can give a contribution to explain the sequences and we opted to insert it in the models we present in this thesis.

1.3 Metric representation of categorical time series

In sequence analysis, one of the main methodological challenges was, and still is, how to suitably measure dissimilarity between two sequences.

The first method presented in literature was Optimal Matching Analysis (OMA). It was originally used in biology to study DNA sequences (Sankoff and Kruskal, 1983) and was afterwards introduced to social science by Abbott and Forrester (1986). OM measures the distance between sequences as the "cost" needed to transform one sequence into another one, using a set of transforming operators: insertion (ι), deletion (δ) and substitution (η). A "cost" is associated to each operator $c(\omega)$, $\omega \in (\iota, \delta, \eta)$. The dissimilarity between two sequences is given by the minimum sum of costs to transform one sequence into the other.

As pointed out by several authors (e.g. Billari, F*ü*rnkranz and Prskawetz (2006); Piccarreta and Billari, 2005) one of the main problems in the application of Optimal Matching concerns the subjectivity of the definition of costs $c(\omega)$. One of the choices applied successfully in literature was to set the insert and deletion cost equal to 1, $c(\iota) = c(\delta) = 1$, and the substitution costs inversely proportional to observed transition frequencies (Rohwer and P*ö*tter, 2004; Piccarreta and Billari, 2005). Given two states s_1 and s_2, let $N_t(s_1)$ be the number of individuals who experiences state s_1 at time t and let $N_{t,t+1}(s_1, s_2)$ be the number of individuals who undergo a transition from state s_1 at time t to s_2 at time $t+1$. The average relative transition frequency from s_1 to s_2 is:

$$f(s_1, s_2) = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{N_{t,t+1}(s_1, s_2)}{N_t(s_1)} \tag{1.3.1}$$

The cost of substituting s_1 to s_2 can then be defined as:

$$c(\eta, s_1, s_2) = 1 - f(s_1, s_2) \quad s_1 \neq s_2 \tag{1.3.2}$$

In some applications it could be reasonable and useful to have a symmetric cost matrix. To this intent the substitution cost between states s_1 and s_2 can be defined as

$$c(\eta, s_1, s_2) = 2 - f(s_1, s_2) - f(s_2, s_1) \quad s_1 \neq s_2. \tag{1.3.3}$$

OM was successfully used in many social science applications (e.g. McVicar and Anyadike-Danes, 2002; Widmer, Levy, Pollien, Hammer and Gauthier, 2003; Stovel and Boland, 2004; Billari and Piccarreta, 2005; Piccarreta and Billari, 2005), but it has also been criticized (Levine, 2000; Wu, 2000; Elzinga, 2003, 2005). Criticisms of OM concern the subjectivity of the choice of the cost determination, the fact that the metric has no social interpretation and that it does not handle duration (the amount of time spent in each status) in a proper way (see Wu, 2000; Levine, 2000; Elzinga, 2003).

Recently, Elzinga (2003, 2005, 2006) developed a series of metrics in the sequence space based on a technique that attempts to solve the drawbacks of the OM technique. The basic idea of the methods, called *combinatorial sequence analysis*, was to use quantified properties of (sub-) set(s) of common subsequences to construct a metric.

Using Elzinga's approach, let $s = s_1 s_2 s_3 ...$ be a sequence of states, s_i belonging to a finite set of possible states. Let $\boldsymbol{S}$ denote the set of all possible sequences of states and $\mathbf{A}$ the set of distinct states; the empty sequence, λ, also belongs to $\boldsymbol{S}$. Given two sequences (s, s'), an attribute is a quantified property $A(s, s')$ of the pair (s, s'). If the attribute satisfies the conditions:

$$\begin{aligned} &A(s, s') = A(s', s) \\ &0 \leq A(s, s') \leq min\{A(s, s), A(s', s')\} \end{aligned} \tag{1.3.4}$$

then the function

$$d(s, s') = A(s, s) + A(s', s') - 2A(s, s') \tag{1.3.5}$$

is a metric over $\boldsymbol{S}$. Several choices are possible to define an attribute, which is then used to define different metrics $d : \boldsymbol{S} \times \boldsymbol{S} \rightarrow \mathbb{Q}$, using (1.3.5).

1.3.1 Elzinga's metrics

In this section we will describe the four metrics Elzinga (2005) provided.

In order to introduce the attributes required to construct these metrics, some concepts and notation are needed. A sequence u is a *subsequence* of s, $u \in s$, if all the states of u appear in s in the same order. Given two sequences (s, s'), we say that u is a common subsequence of s and s' if $u \in s$ and $u \in s'$, i.e. $u \in C(s, s')$, where $C(s, s')$ denotes the set of all common subsequences of (s, s'). It is possible that a subsequence is embedded more than once in a sequence. If a subsequence u is embedded k times in a sequence s we write $|s|_u = k$. A sequence s is said to be of length n if it is constructed with n, not necessarily distinct, characters from $\mathbf{A}$ and this fact is denoted by writing $|s| = n$. The subsequence $s^i = (s_1 ... s_i) \in S$ is called the i-th prefix of s with $s^0 = \lambda$ and $s^n = s$ if $|s| = n$.

The possible attributes are:

1. The Longest Common Prefix ($LLCP$):

$$LLCP(s, s') = max_i s^i : s^i = s'^i \tag{1.3.6}$$

   denoting the length of the longest common prefix of the pair (s, s'). The related Euclidean distance is

$$d_{LLCP}(s, s') = |s| + |s'| - 2LLCP(s, s') \tag{1.3.7}$$

2. The Length of the Longest Common Subsequence ($LLCS$):

$$LLCS(s, s') = max_{u \in C(s,s')} |u| \tag{1.3.8}$$

   denoting the length of the longest subsequence of the pair (s, s'). The related Euclidean distance is

$$d_{LLCS}(s, s') = |s| + |s'| - 2LLCS(s, s') \tag{1.3.9}$$

3. The Number of Common Subsequences (NCS):

$$NCS(s, s') = |C(s, s')| \tag{1.3.10}$$

and the related Euclidean distance is

$$d_{NCS}(s, s') = NCS(s, s) + NCS(s', s') - 2NCS(s, s') \tag{1.3.11}$$

4. The Number of Matching Subsequences (NMS):

$$NMS(s, s') = \sum_{u \in C(s,s')} |s|_u |s'|_u \tag{1.3.12}$$

which weighs the common subsequences by their embedding frequency in both sequences. The related Euclidean distance is

$$d_{NMS}(s, s') = NMS(s, s) + NMS(s', s') - 2NMS(s, s') \tag{1.3.13}$$

Elzinga also suggested criteria to take into account the duration of the permanence in each status: minimal shared time and duration as weights.

The first approach that incorporates the duration in attributes considers the part of the duration that is the same in the common subsequences of two trajectories. Let $u \in (s, s')$ and let $t_s(u_i)$ denote the time spent in the i-th state of u as embedded in s. The minimal shared time $t(u)$ is defined as

$$t(u) = \sum_{i=1}^{|s|} min(t_s(u_i), t_{s'}(u_i)) \tag{1.3.14}$$

and the distance is calculated by weighing the common subsequences or their embedding by the minimum shared time. In the second approach, the time spent in a common subsequence is defined as a vector product of state duration vectors:

$$t(u) = \sum_{i=1}^{|s|} t_s(u_i) t_{s'}(u_i) \tag{1.3.15}$$

The main difference between the two approaches is that "duration as weight" for each pair of common subsequences considers both the duration of the subsequence in two trajectories, while "minimal shared time" is insensitive to the longer of the two compared durations.

All the distance matrices in this thesis are calculated with the CHESA 2.1 package. For details of this software see *http://home.fsw.vu.nl/ch.elzinga/*.

It is not possible to show that one of these metrics is better than the others and there is no criterion to lead to a permanent rank of them. Only social science theory, the accurate study of the aspects that each metric emphasizes and the interpretation of the results obtained can help choose which metric is more appropriate for the subject in analysis. Here, the goal is to select the best metric in partitioning sequences, in the sense that, as input of cluster analysis, provides clear ideal type trajectories and highlights the main difference among sequences.

A limit of OM and Elzinga's dissimilarities (except the $LLCP$) is that they do not take into account the direction of the sequences, that is, when they evaluate equal subsequences (e.g. NCS) or a different state (OM) between two sequences they do not consider the position of such subsequences or state in the sequences. For example, it is different if two sequences have a common subsequence at the beginning of the sequence or if one has it at the beginning and the other one at the end of the sequence. NCS counts the number of common subsequences, but does not take into account if the subsequences refer to similar or dissimilar life periods. It is also different if two sequence differ for a state in a slightly different order (for example AAAAB and AAABA) or in in completely opposite position (AAAAB and BAAAA), but OM does not pick up this

difference. Probably it could be useful to introduce opportune weights to account for the position of a states when different states or common subsequences are evaluated. To propose a new metric is beyond the scope of this thesis, but it is important to note that the direction of the sequence has a significant role.

1.4 Cluster and MDS: a graphical tool

Cluster analysis is used to segment a collection of cases into homogeneous groups according to a dissimilarity measure. Given the dissimilarity matrix standard techniques can be applied to obtain clusters (McVicar and Anyadike-Danes, 2002, discuss criteria of clustering). We use a hierarchical cluster analysis as in McVicar and Anyadike-Danes. We use the Ward's agglomerative algorithm as in Aassve, Billari, Piccarreta (2004) and Piccarreta, Billari (2007), because it generally produces a set of clusters of more homogeneous size as compared to other common algorithms (e.g., single linkage, complete linkage, centroid, median).

It is well known that in cluster analysis different measures of distance can be used (e.g., statistical distance, Manhattan distance, a.s.o.). Usually, the proper distance measure is selected by referring to the quality and the meaningfulness of the obtained clusters. As concerns the quality of a partition, this can be evaluated by considering the homogeneity within clusters. As for the meaningfulness of clusters, this is related to the interpretability and to the separation of clusters with respect to the variables at the basis of the agglomerative procedure. In sequence analysis the "objects" to be clustered are very complex, and it is not very easy to evaluate in depth clusters' characteristics. To overcome this problem we follow the procedure proposed by Lior, Piccarreta (2007). It refers to Multidimensional Scaling (see e.g. Lattin, Green, Douglas Carroll, 2002), a statistical technique which may prove useful to obtain a graphical representation of

the sequences (within clusters) and to clarify which are the main differences between sequences emphasized by a given dissimilarity measure. The input in Multidimensional Scaling (MDS) is a dissimilarity matrix. Each case is projected onto a factorial space of properly chosen dimension. The extracted dimensions may be interpreted as the latent factors underlying the observed dissimilarities between cases. More precisely, the Euclidean distances calculated by referring to the factorial dimensions should be as close as possible to the observed dissimilarities. A real vector is associated to each sequence. If the projection space has a low dimension, it is possible to plot the original sequence-points in a vectorial space, and moreover, to deal with low-dimension vectors rather than with the whole sequences. Halphin and Chan (1998) represented the MDS solution applied to sequence analysis in a three-dimensional scatterplot in order to investigate the coherence and the interpretability of the space in which the sequences were plotted. We move the attention from the description of the MDS space to the description of the sequence features emphasized by the metric used. We thus do not focus on the interpretability of the MDS space but on the characteristics that make two sequences similar according to a given metric and on the interpretation of the clusters obtained with that metric. After having applied MDS, the procedure can be described as follows:

1. order sequences according to the first MDS factor. This permits analyzing the characteristics of the sequences that are evaluated as more similar/dissimilar on the basis of a given metric;

2. plot the sequences in the mds factorial space so as to visualize which sequences are clustered together by a hierarchical algorithm based on a given metric.

In this way, it can be possible to compare the different metrics described in the previous section and to choose the one providing a suitable MDS solution and/or clustering. However, as Elzinga (2006) remarks, we are aware that this does not mean that the chosen metric is superior to any other metric, but that it underlights the more meaningful aspects of the sequence regarding the purpose of the analysis.

In our analysis we took into account only metrics which, at least in our opinion, use information in sequences in the most complete manner: NCS and NMS, handling duration as a weight and minimal shared time, and OM, using symmetric and not symmetric data driven cost matrices. Hence, we excluded $LLCP$ and $LLCS$ because they are based upon a partition of the sequence only (the prefix or subsequence) and hence partially use the information contained in the sequences. We compared the metrics through some preliminary analyses that consider the order of the sequence according the first MDS factor and the representation of clusters in the Multi Dimensional Scaling space (results not shown here). Looking at these graphics we noted that in considering two sequence similar, the OM metric emphasizes the dominant state, while NCS and NMS stress more the life-path. The first thus is more focus on the "broad" vision of the sequence, unlike NCS and NMS that consider all the visited states, regardless of the time spent in them. For example, OM considers two sequences as S/144 and S/100 U/2 S/42 quite close, on the contrary of the other two metrics. From a theoretical point of view, we believe that in family formation field it is important to take into account all the state changes, since also small changes can influence the life choice of individuals. Therefore we prefer the NCS and NMS metrics. This preference is confirmed also from a "practical" point of view, evaluating the quality of partitions. Looking at the results of cluster analysis we decided to use the NMS metric with duration as weight to obtain the distance matrix that we will use as a starting point for the methods to be applied in this and the next chapters. Clusters obtained

with NMS are well separated and could be easily described in terms of dominance of states in their life history (see below), while clusters obtained with other metrics result separated but can not be so easily interpreted in terms of sequence patterns (see the Appendix).

As it was said before, a meaningful graphical representation of a sequence data matrix is a basic, but not trivial matter. At this aim we display the sequences, ordered according to multidimensional scaling coordinates, in a plot having time in months on the vertical axis and the rank of individuals on the horizontal axis. Different colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC (refer to section 1.2 for definitions). This sorting, bringing similar sequences closer, allows the observation of the main trends of the life courses based on the color impact. Figure 1.2 shows the sequences in our sample. Such a representation remains difficult to interpret but it is possible to note the second and third states are crucial in determining the similarity between sequences. Moreover, it is possible to note that this metric distinguishes between two main typologies of family formation: single-marriage-marriage with children (S M MC), on the left side of the graph, and single-cohabitation (S U), on the right side of the graph.

Certainly, deeper analyses are required to investigate this complex set of relationships, but this kind of representation is however informative. It will allow us to evaluate graphically the results of the more detailed studies which will be described in this chapter and in the next one.

A common approach to exploratory sequence analysis has been to use dissimilarity data to perform cluster analysis in order to find ideal type trajectories (Abbott, 1995; Billari, Piccarreta, 2001, 2005; McVicar and Anyadike-Danes, 2002; Mouw, 2004, Elzinga and Liefbroer, 2006). The data considered here were grouped using a hierarchical cluster analysis with Ward's algorithm. The combination of "Cluster-MDS" graphics is a good

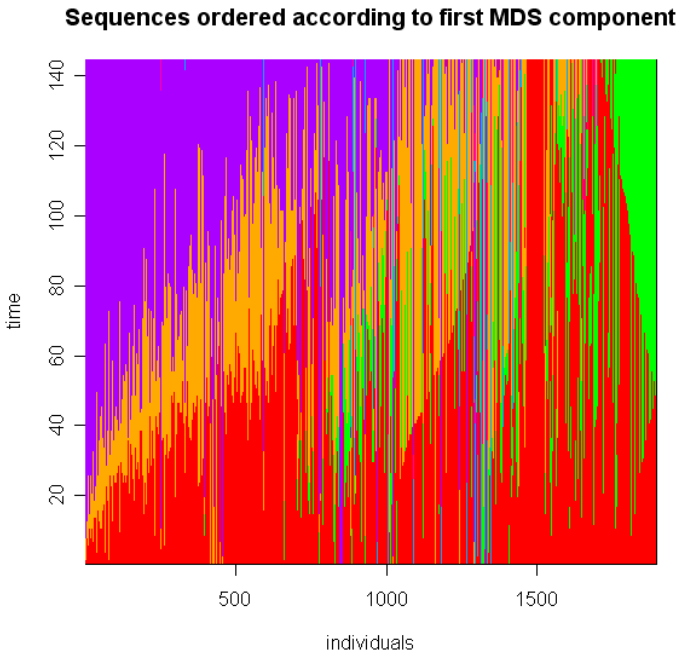

Figure 1.2: Representation of individual sequences of states using first MDS component ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC).

instrument to characterise and explain the clusters.

Analysing the "Cluster-MDS" graphics obtained by varying the numbers of groups we opt for 6 groups. The contents of the clusters are homogeneous and the cluster are distinct, when looked at as "family careers". NMS and clustering lead to a reasonable typology of family life trajectories. For brevity, we report only the plots based on this choice. In Figures 1.3-1.4 we plot together the clusters and the MDS solution, coloring the points of MDS according to cluster membership (black = cluster 1, red = cluster 2, green = cluster 3, blue = cluster 4, sky-blue = cluster 5, pink = cluster 6) and representing the sequences, separated in clusters and ordered by the MDS coordinates. In order to deal with the fact that the sizes of the groups are often very different, we scale the width of the sequences so as to have the same overall span in the horizontal

axis. This allows for a better appreciation of the difference and the similarities in the groups.

The clusters obtained are generally well defined and display significantly distinct pat-

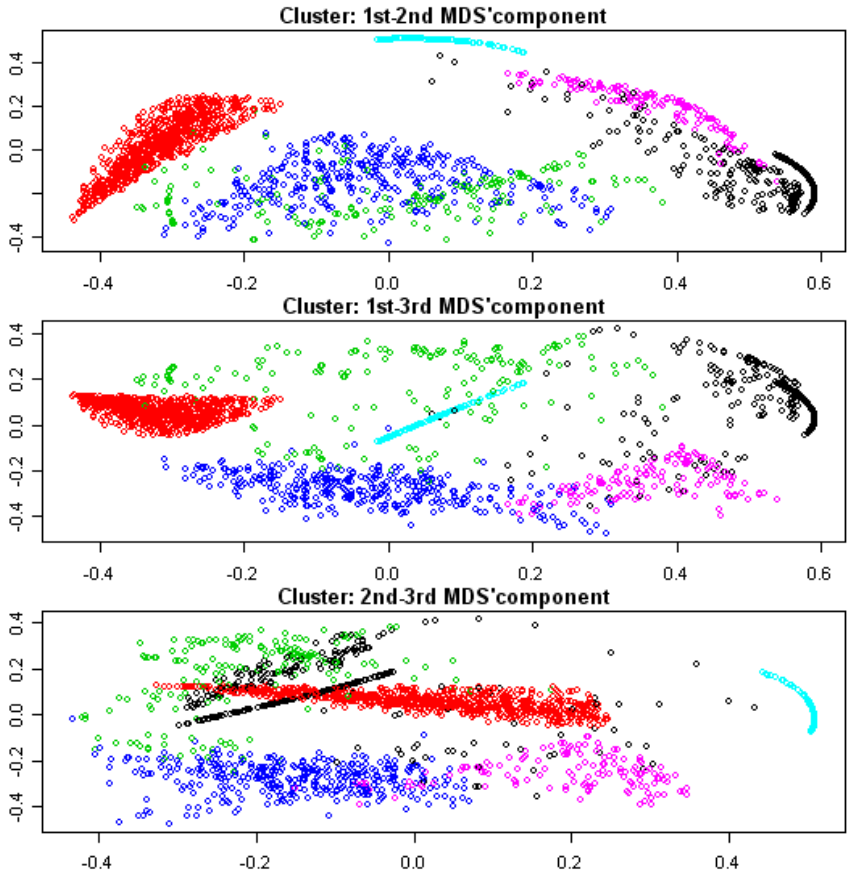

Figure 1.3: Projection of the three-dimensional multidimensional scaling solution on the xy, xz and yz planes. Points are colored according to the cluster membership: black = cluster 1, red = cluster 2, green = cluster 3, blue = cluster 4, sky-blue = cluster 5, pink = cluster 6.

terns, briefly described Table 1.1. The first cluster is dominated by single-cohabitation (S U) combinations. This cluster also include individuals that experience only the single state (S). In the second cluster there are only individuals that experience the single-marriage-marriage with children dynamic; the time spent in each state may vary

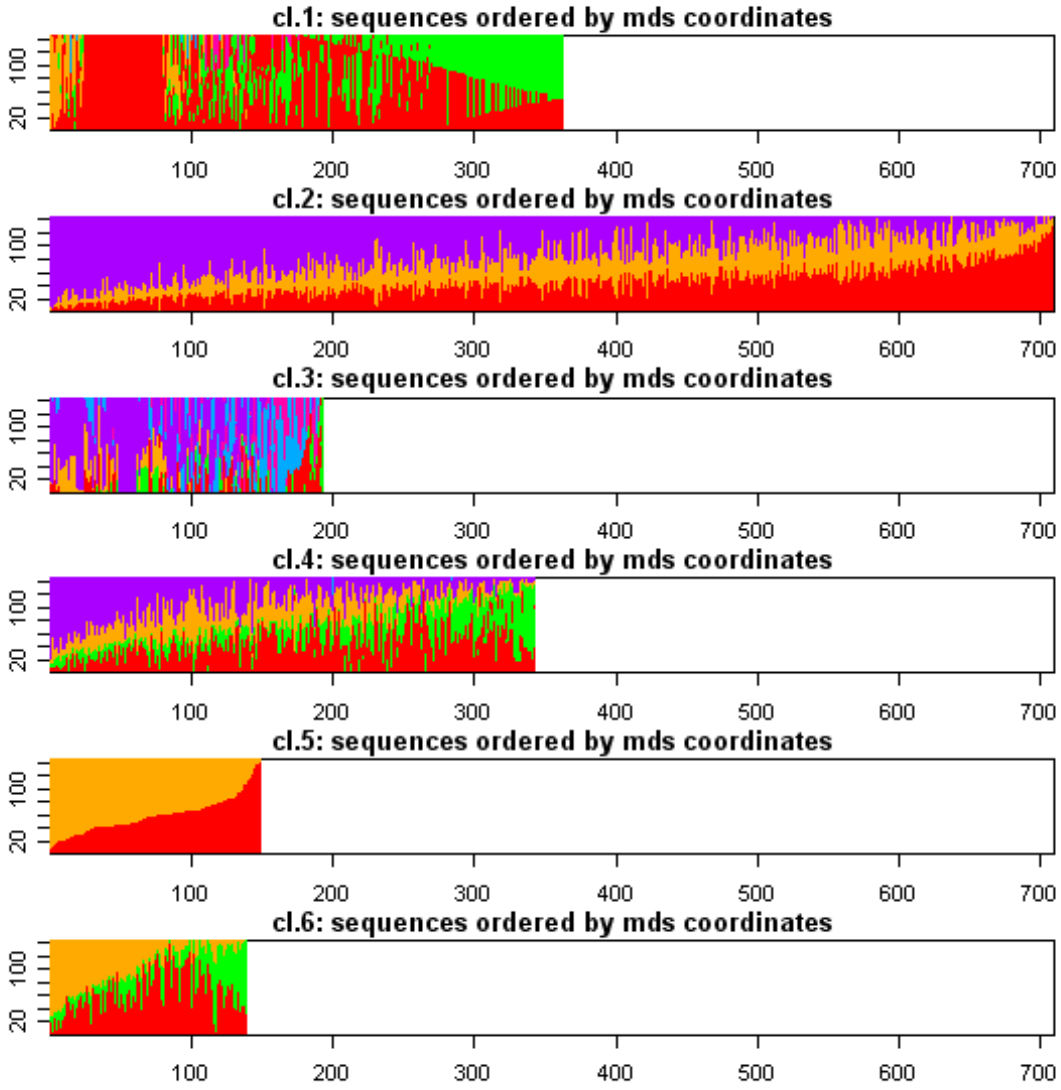

Figure 1.4: Representation of individual sequences of states, separated in clusters, using first MDS component ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC).

(S M MC). Cluster 3 is a residual cluster containing the less frequent patterns, such as complex sequences (with many transition) or marriage with children throughout

Cluster	N	Brief description
1	363	single-cohabitation (S U) sequences
2	709	single-marriage-marriage with children (S M MC)
3	193	residual category
4	343	single-cohabitation-marriage-marriage with children (S U M MC) sequences
5	150	single-marriage sequences (S M)
6	139	single-cohabitation-marriage (S U M) sequences

Table 1.1: Description for the obtained cluster solution

the observation period. The fourth cluster is dominated by the single-cohabitation-marriage-marriage with children (S U M MC) sequences. Cluster 5 contains single-marriage sequences (S M). Cluster 6 is composed of single-cohabitation-marriage (S U M) sequences. In the last two clusters the time spent in each state varies flatly.
Figures 1.5-1.11 show that the sequences are strongly structured looking at the number of states experienced and the permanence in different states. We plot the first two dimensions of the MDS solution using different colors to represent the different features of the sequences. Colors for the number of states visited are: black = 1, red = 2, green = 3, blue = 4, sky-blue = 5, pink = 6, yellow = 7, grey = 8, violet = 9. For the time spent in different states we divide the distribution of time in quartiles and color the first quartile in black, the second in red, the third in green and the fourth in blue. Figures concerning the time spent in SC and UC are less structured because only few women experience these two states.

These preliminary analyses confirm that NMS metric exploits in a complete way the informations in sequences in order to distinguish individual behavior: the number of states experienced, the second and third states experienced (the first state is not particularly informative since it is almost invariably S) and time spent in different states.
To sum up, the combination of cluster analysis and Multidimensional scaling provide an appropriate tool: to choose the metric and to represent the sequences in a simple

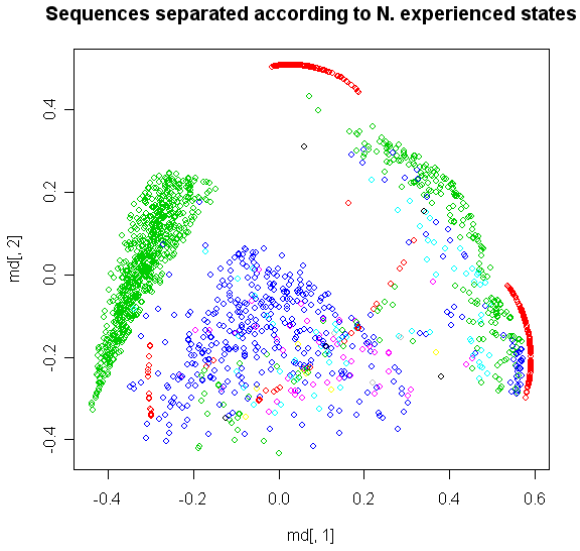

Figure 1.5: Projection of multidimensional scaling solution. Points are colored according to the number of states visited by individuals: black = 1, red = 2, green = 3, blue = 4, sky-blue = 5, pink = 6, yellow = 7, grey = 8, violet = 9.

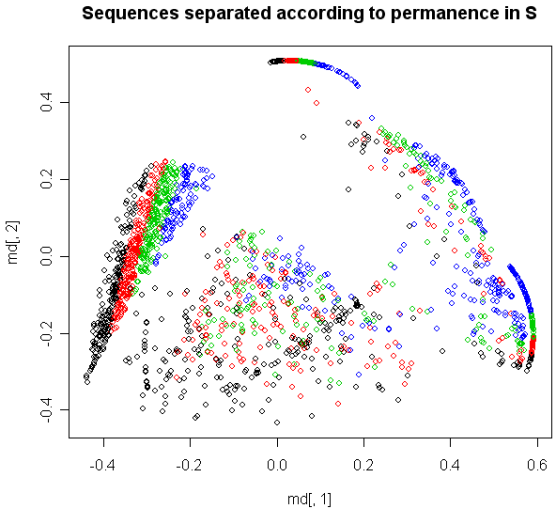

Figure 1.6: Projection of multidimensional scaling solution. Points are colored according to the quartiles of the distribution of the time spent by individuals in state "single" (black = 1^{st} quartile, red = 2^{nd} quartile, green = 3^{rd} quartile, blue = 4^{th} quartile).

and easy to interpret manner.

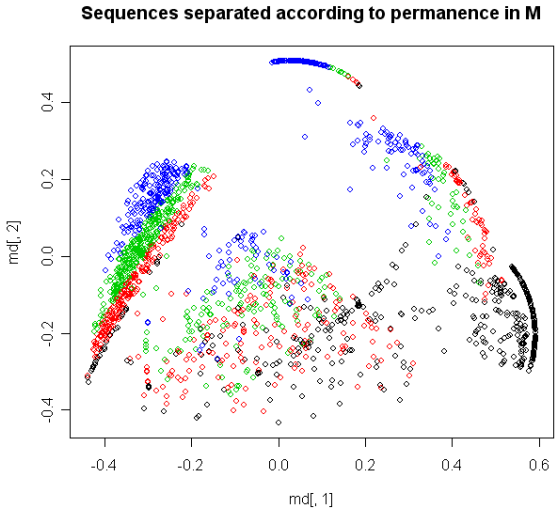

Figure 1.7: Projection of multidimensional scaling solution. Points are colored according to the quartiles of the distribution of the time spent by individuals in state "marriage" (black = 1^{st} quartile, red = 2^{nd} quartile, green = 3^{rd} quartile, blue = 4^{th} quartile).

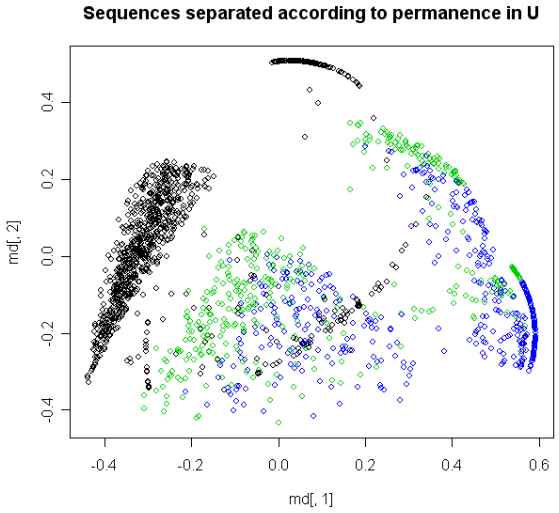

Figure 1.8: Projection of multidimensional scaling solution. Points are colored according to the quartiles of the distribution of the time spent by individuals in state "cohabitation" (black = 1^{st} quartile, red = 2^{nd} quartile, green = 3^{rd} quartile, blue = 4^{th} quartile).

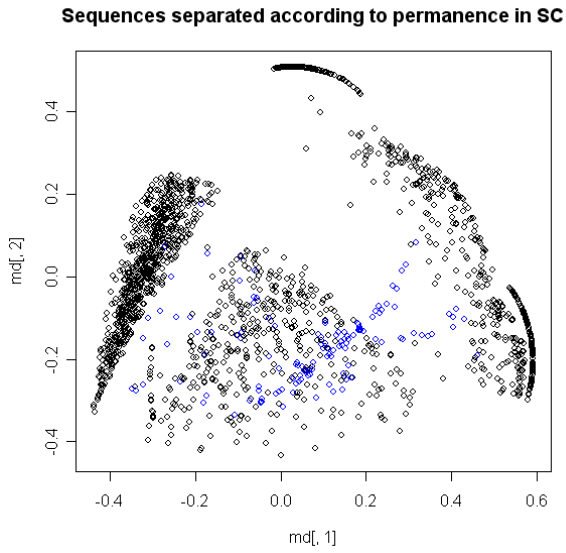

Figure 1.9: Projection of multidimensional scaling solution. Points are colored according to the quartiles of the distribution of the time spent by individuals in state "single with children" (black = 1^{st} quartile, red = 2^{nd} quartile, green = 3^{rd} quartile, blue = 4^{th} quartile).

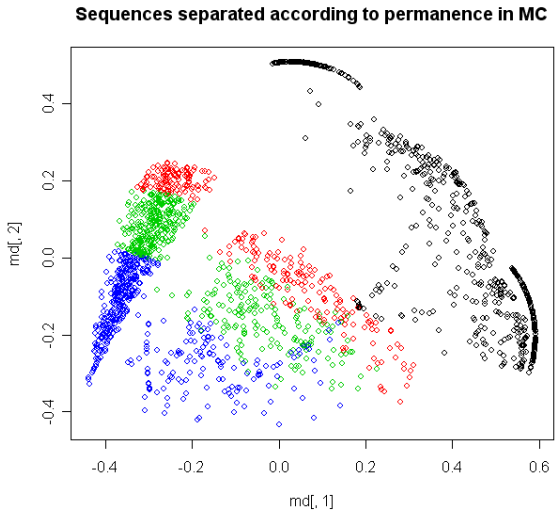

Figure 1.10: Projection of multidimensional scaling solution. Points are colored according to the quartiles of the distribution of the time spent by individuals in state "marriage with children" (black = 1^{st} quartile, red = 2^{nd} quartile, green = 3^{rd} quartile, blue = 4^{th} quartile).

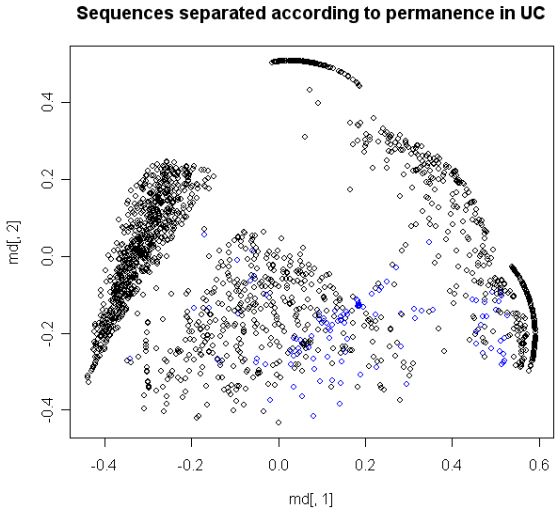

Figure 1.11: Projection of multidimensional scaling solution. Points are colored according to the quartiles of the distribution of the time spent by individuals in state "cohabitation with children" (black = 1^{st} quartile, red = 2^{nd} quartile, green = 3^{rd} quartile, blue = 4^{th} quartile).

1.5 Explaining and predicting life courses: the methods

A criticism of existing approaches to sequence analysis is that the results of cluster procedure have not generally been applied to further explanatory analysis with success. McVicar and Anyadike-Danes (2002) used the patterns found as the outcome of cluster analysis as the input for a multinomial model in order to predict sequence patterns using covariates. We replicate the method proposed by McVicar and Anyadike-Danes for Dutch FFS data, using clusters as a dependent variable in a multinomial logit model to understand if individual and family characteristics (education, religion, parental divorce, cohort) have an effect on the probability of experiencing a specific sequence pattern. Refer to Section 2 for the description of these variables. Evidence suggests that these background characteristics influence the probability of experiencing a specific pattern of family status. In Table 1.2, we describe the results of the model. It reports

Covariate	Coefficient for the following comparisons				
	Cl.1 vs Cl.6	Cl.2 vs Cl.56	Cl.3 vs Cl.6	Cl.4 vs Cl.6	Cl.5 vs Cl.6
Cohort	**0.54**	**-1.15**	**-0.93**	**-0.05**	**-0.99**
Divorce	0.61	-0.72	**0.19**	0.47	0.10
Religion	-0.14	**1.13**	-0.04	0.03	**1.12**
Education	-0.08	**-1.31**	**-1.63**	**-0.63**	**-0.73**
Intercept	**2.14**	**4.76**	**4.33**	**2.00**	**2.17**

Table 1.2: Logit model coefficients for the six-clusters solution (Boldface indicates p-values lower than 0.05).

the estimated coefficients of the logit model. Although all the covariates are significant according to the Wald Chi-Square test, once we try to predict cluster membership on the basis of the multinomial model, we observe a very low predictability. This is illustrated by Figures 1.12-1.13. These figures compare the obtained clusters with those actually predicted using the multinomial model. It is immediately noticeable that two clusters are never predicted in the prediction; moreover the largest sample cluster is overestimated by the model and the original subdivision is not respected in the predicted groups, which end up resulting not homogeneous.

From our point of view, the principal limitation of this method is due to the loss of information that follows the grouping of sequences in clusters. The transformation from a high dimensional structure into a one-dimensional categorical variable presumably does not capture enough information to make a reliable forecast. Cluster analysis is a useful tool to simplify the sequences structure in order to identify patterns, but it does not seem suitable for prediction. We believe that a more successful approach to explain and predict the structure of data is to deal with less simplified data. For example, one could consider the distance between sequences, or retain the full complexity of the sequences. A dissimilarity matrix is however a lower-dimensional summary of

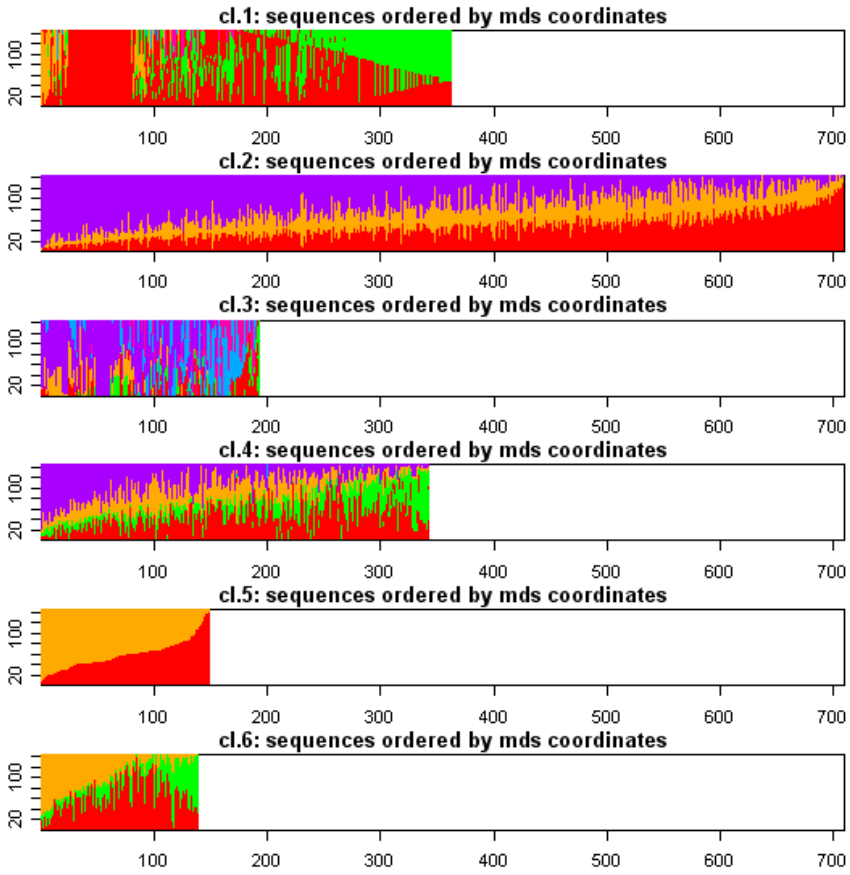

Figure 1.12: Representation of individual sequences of states, separated in clusters, using MDS ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC).

the sequences and clearly it is not a "sufficient statistic" in any sense, but it is more informative than a cluster variable. In what follows, we propose three methods in these directions:

1. a regression model with distance between sequences and a suitably chosen template as the dependent variable;

2. an ANOVA-like model building approach based on permutation distribution ideas;

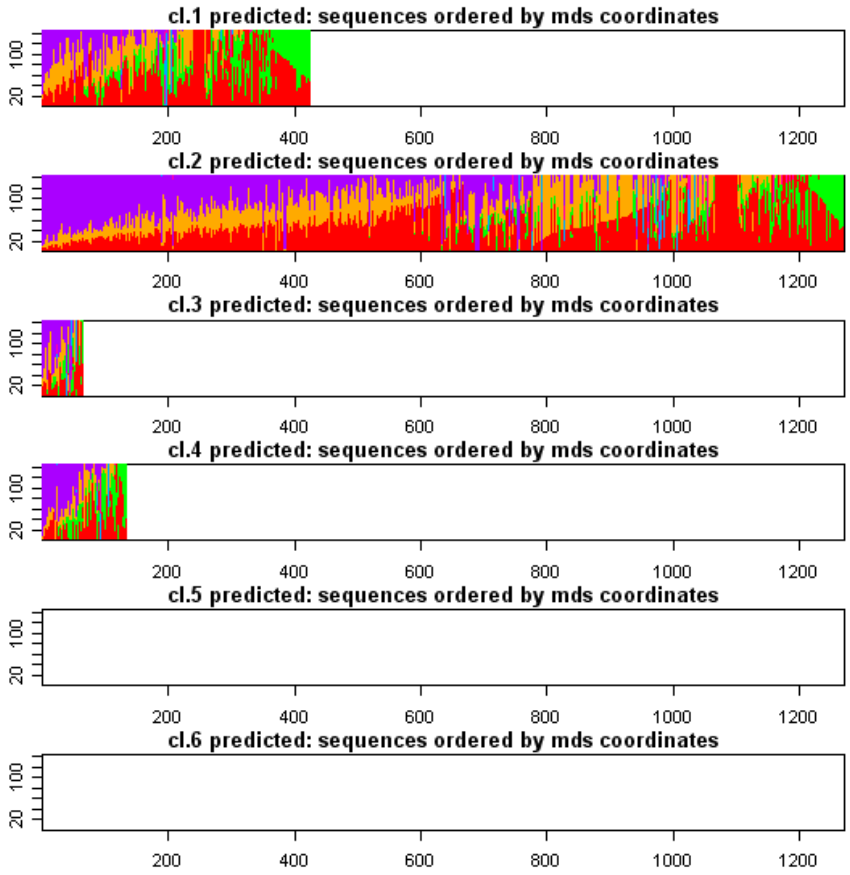

Figure 1.13: Representation of individual sequences of states, separated according to estimated clusters, using MDS ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC).

3. a parametric approach to model the whole process that generates sequences.

We now briefly describe the content of the chapters that follow.

In Chapter 2, we present two regression models to analyse adolescent premarital pregnancy. The goal is to seek individual and origin family factors associated with the propensity to be an adolescent single mother. Moreover, we emphasize the importance of selecting the more appropriate metric with regard to the issue in analysis.

In Chapter 3, we discuss and apply Analysis of Dispersion (ANODI) and selection

model techniques to investigate the determinants of the distance between sequences. The method allows to analyse the explanatory capability of the factors and to evaluate their significance using permutation distribution ideas.

In Chapter 4, following a parametric approach we model the whole process that generates life trajectories with a combination of time-to-event distributions and transition probabilities. With the model it is possible to estimate the transition probabilities and the duration distributions between subsequent transitions, as well as to compute the probability that a given individual experiences a certain transition status. Covariates can also explain differences between groups of individuals.

1.6 Appendix

In this Appendix we show the results of cluster analysis using NCS (with duration as weights) and OM metrics (with symmetric cost matrix).

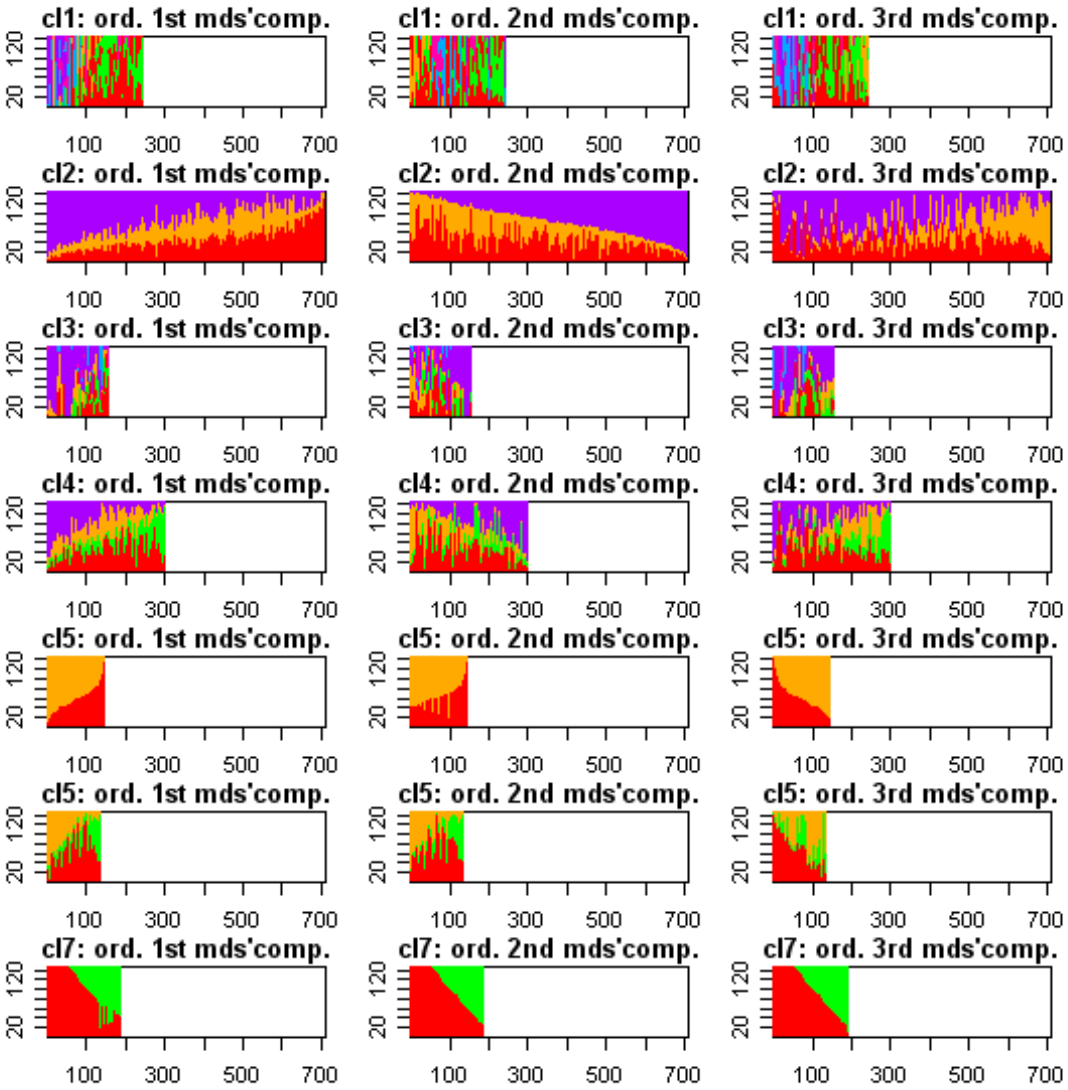

Figure 1.14: Representation of individual sequences of states, separated in clusters, using first three MDS component ordering (NCS with duration as weights). Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC).

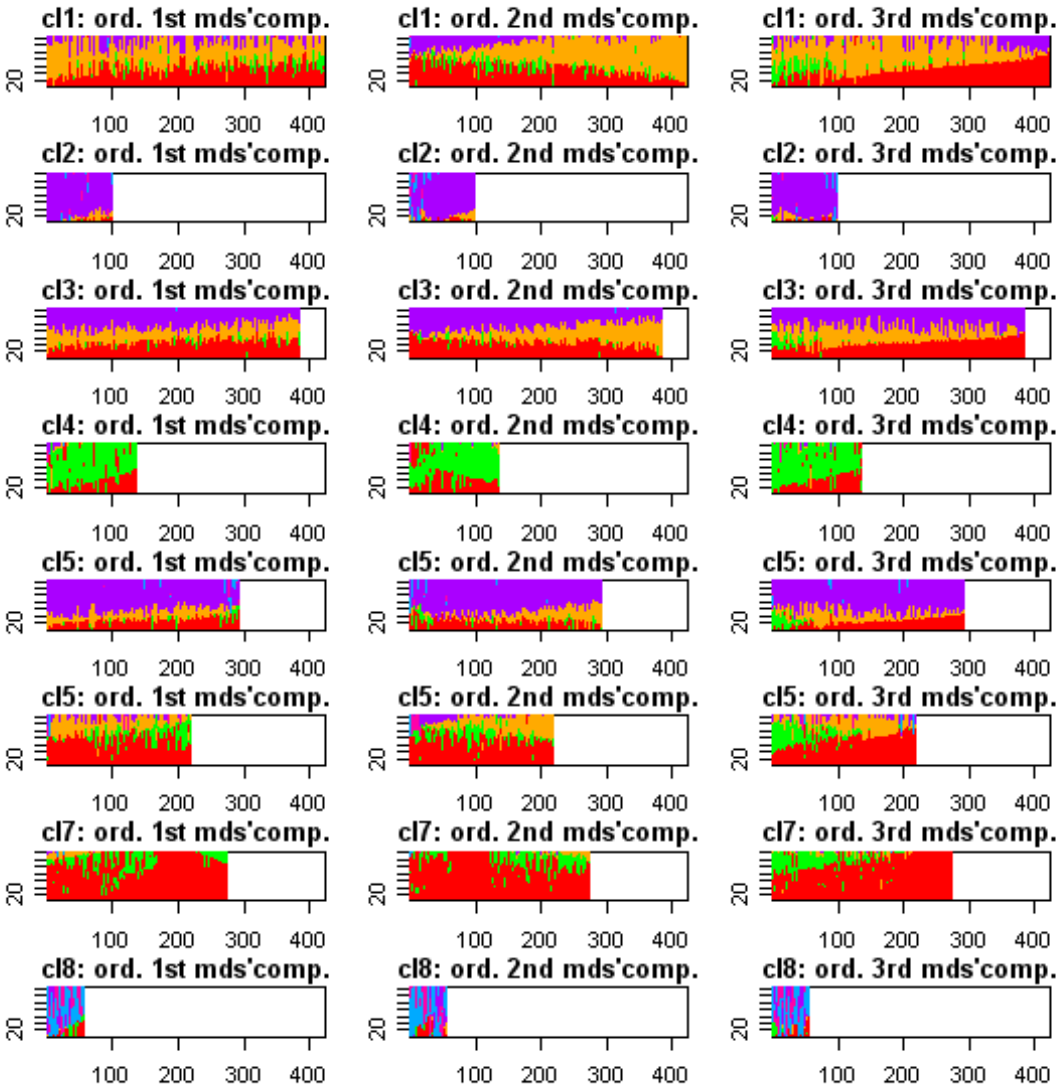

Figure 1.15: Representation of individual sequences of states, separated in clusters, using first three MDS component ordering (OM with with symmetric cost matrix). Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC).

Chapter 2

Predicting teenage and long run non-marital childbearing using sequence based methods

2.1 Introduction

Identifying young women who are likely to become early single mothers is an interesting issue for policy makers, because the prevention of "risky" behavior might be possible with targeted and well-timed intervention. Evidence suggests that teenage premarital motherhood is associated with a number of negative outcomes for the mother and the child. For the mother: worse life opportunities, educational underachievement, prolonged welfare dependence and maternal depression. For the child: learning problems, delinquency and addiction (Hofferth, 1987; Bardone, Moffitt, Caspi, Dickson and Silva, 1996; Moore, Morrison and Green, 1997). The development of models that use antecedent and current life factors to explain the risk of teenage pregnancy has an enormous importance in identifying targets for future intervention and improving the success of programs that prevent this problematic behavior.

In this paper we use sequence techniques to predict the predisposition of being a long-term single mother. We adopt an international comparative approach, and we study nine European and North American countries for which comparative retrospective data

on union and birth histories have been collected in Fertility and Family Surveys. The only attempt in the literature to detect and predict sub-groups as potential targets of specific policies using sequence analysisis from McVicar and Anyadike-Danes (2002). They extracted some ideal-type trajectories with cluster analysis and tried to predict cluster membership using several socio-economic variables through a logit model. Nevertheless their method, applied to other data, performs quite poorly as far as forecasting is concerned (see Chapter 1). We believe that the principal limitation of this method is due to data loss following the grouping of sequences in clusters. The transformation from a high dimensional structure into a one dimensional categorical variable presumably does not capture enough information to make forecasting reliable. We propose a transformation of the multidimensional sequence into a one dimension quantitative variable (a distance variable), so that it remains easily predictable in a regression model, while containing more information than cluster variable. The approach we adopt is to define a template, representing the long-term single mother pattern, i.e. a women who has had children and has been single during the entire observation period (age 18-30). Then we calculate the distance, transformed to the open unit interval (0,1), between this template and the observed sequence for each woman in our sample. The bigger the distance, the smaller the proximity to the single mother pattern. The distance obtained was used as the dependent variable in a regression model to identify which factors contribute to predict the proneness of being a long-term single mother.

Note that the method allows to utilize templates not in the sample and in our application this is the case. This is particularly useful when the object in analysis concerns not frequent events and the available datasets contain an insufficient number of cases to apply other statistical techniques.

The first aim of this work is to explain which background characteristics, that have

been correlated with early pregnancy in literature, affect the distance from the problematic single mother trajectory. In order to do this we consider education level, cohort, religiousness and family background.

There is an abundances of literature showing that education is a factor that increases the motivation to control fertility. Jessor and Jessor (1977), Hanson, Myers and Ginsburg (1987), Plotnick and Butler (1991), Plotnick (1992) and Tanfer, Cubbins and Brewster (1992) found evidence that teenagers with positive attitudes toward school and with high long run educational aspirations have lower propensity to premarital childbearing, because they generally pursue long-term goals that make them more careful contraceptive users and more oriented to abortion in case of unwanted pregnancy.

Belonging to a younger cohort should also have the effect of reducing the risk of single childbearing due to increased availability of modern contraceptives, acceptance of abortion and the growing participation of women in higher education and professional activities.

Although most religions proscribe premarital intercourse and thus religious women may have lower propensity to become pregnant before marriage (DeLamater, 1981; Thornton and Camburr, 1989), proscriptions against contraception and abortion have effects in the opposite direction and thus religiousness is potentially a factor that increases the propensity of premarital pregnancy (Plotnick, 1992; Tanfer et al.,1992).

Certain family background characteristics are known to influence fertility and family formation (Micheal and Tuma, 1985) and, in particular, premarital pregnancy and its resolution (Hoffert and Hayes, 1987; Miller and Moore, 1990). Several studies have highlighted that a family background with a single mother or multiparental transitions as a result of marital breakdown can lead to earlier initialization to sexual intercourse and higher propensity of a teenage pregnancy in US (Capaldi, Crosby and Stoolmiller,

1996; O'Connor, Thorpe, Dunn and Golding, 1999; Wu and Martinson, 1993; Woodward, Fergusson and Horwood, 2001). Moreover Howard and Powell (2002) found, using US data, that family structure influences contraceptive decisions. Women raised by both parents from birth to age 14 are likely to use more effective methods of contraception and for this reason are less exposed to the risk of an unwanted pregnancy.
The second aim of this work is to highlight the necessity of choosing an appropriate metric as the starting point for successive analysis. We use different metrics to obtain distances and then we compare the results. We do not attempt to evaluate the metrics as to their general appropriateness, but to select the most suitable metric with regard to the subject of the analysis.
The remainder of this paper is organized as follows. Section 2 provides a brief description of the data and methods. In Section 3 we present our main results. In Section 4 we formulate conclusion and comments.

2.2 Data and methods

The next subsections describe the data, the independent variables in the model and the methods we use for our analysis.

2.2.1 Data

Childbearing outside marriage is a life course event in which European states exhibit a high heterogeneity (Kiernan, 1999), that is related to cultural and historical patterns. We examine the childbearing outside of a stable union according to background characteristics of the women in states characterized by different historical traditions and institutional frameworks: Italy, Spain, the Netherlands, Poland, Estonia, Sweden, Finland, USA and Canada. The data for this analysis are from the Fertility and Family Survey (FFS) described in detail in Chapter 1.

The dependent variable is a transformation of the distance between each trajectory in the sample and the template SC/145. Due to the presence of large distances in a number of cases for the NMS distance, following the transformation to the interval $(0, 1)$, almost all the distances are close to 0. As a consequence, the results may not be meaningful because of the absence of enough variance in the dependent variable. For this reason the observations which lead to extreme distances were removed: 11 sequences for Netherlands (0.006 of the country sample), 4 sequences for Estonia (0.006 of the country sample), 6 sequences for Sweden (0.005 of the country sample), 4 sequences for Finland (0.004 of the country sample), 52 sequences for USA (0.012 of the country sample) and 6 sequences for Canada (0.002 of the country sample). The removed individuals are characterized by particular and very complex sequences, for the considerable number of transitions and/or of distinct states, and thus not strictly connected with the phenomenon being studied.

2.2.2 Explanatory variables

The explanatory variables [1] included in the model are:

1. The level of education. Using level 1 as baseline category in the regression model the other two levels are inserted in the analysis as two dummy variables (Educ2, Educ3)
2. Religiousness (Religion)
3. Parental divorce (Divorce)
4. Cohort (Cohort)

Table 2.1 gives the covariate and sample size in each country.

[1] For a detailed description of the covariates see Section 1.2

	COVARIATES									
Country		Educ		Religion		Divorce		Cohort		Sample size
	1	2	3	No	Yes	No	Yes	53-57	58-62	
Sweden	0.11	0.19	0.70	-	-	0.89	0.11	0.49	0.51	1316
Finland	0.14	0.16	0.70	-	-	0.94	0.06	0.73	0.27	1065
USA	0.50	0.01	0.49	0.08	0.92	0.81	0.19	0.49	0.51	4186
Canada	-	-	-	0.02	0.98	0.94	0.06	0.54	0.46	2672
Poland	0.08	0.28	0.64	0.01	0.99	0.97	0.03	0.59	0.41	1313
Estonia	0.11	0.0.53	0.36	0.51	0.49	0.84	0.16	0.50	0.50	651
Netherlands	0.09	0.37	0.53	0.37	0.63	0.93	0.07	0.48	0.52	1886
Italy	0.32	0.17	0.51	0.08	0.92	0.98	0.02	0.50	0.50	1559
Spain	0.48	0.20	0.32	0.17	0.83	0.98	0.02	0.47	0.53	1320

Table 2.1: Sample size for each category;
Educ=1: no education after age 15
Educ=2: 0-3 years of education after age 15
Educ=3: 3+ years of education after age 15

2.2.3 Methods

We consider the propensity to become a teenage single mother as the inverse of the distance between the sequence that represents the life course of a woman in the sample, and a template which represents the teenage single mother pattern. Thus we define the template as the sequence Single with at least one child for 145 months, between the ages 18 to 30 (SC/145). We compute the distance d with three different metrics described in Chapter 1. In most studies on sequence analysis, the distance matrix is considered as a given starting point, but the choice of the metric to measure the distance between two sequences is not a trivial question and has consequences for the results. This is mainly due to the fact that different metrics can reverse the order of distances. In this paper we therefore compare three different metrics: Optimal Matching (OM), the metrics based on the Length of the Longest Common Subsequence (LLCS) and the Number of Matching Subsequence (NMS). We then transform the distance to unit interval [0,1], by taking $d' = (d - m)/(M - m)$, where m and M are the minimum and the maximum

distance in the sample. The beta regression and fractional logit models are suitable to study the effect of the covariates on this limited dependent variable. From the shapes of the empirical distance distributions, we conclude that the assumption of normally distributed distance is not acceptable (see also Figure 2.2).

The beta regression model is directly related to an extended generalized linear models framework for joint modeling means and dispersions described in McCullagh and Nelder (1989, Ch. 10). It assumes that the dependent variable is distributed according to a Beta distribution, which provides for a rich class of distance distributions (Figure 2.1 display several examples). Moreover it permits the modeling of both mean and variance, including modeling heteroskedasticity, through a reparametrization of the distribution. In the conventional parametrization

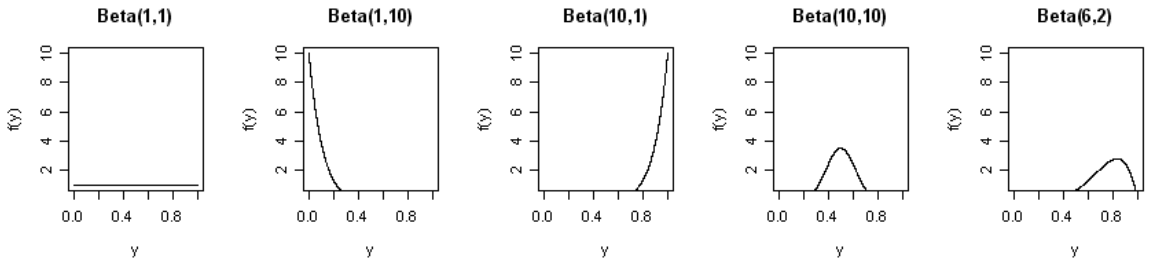

Figure 2.1: Beta densities for different combinations of (α, β).

$$f(y|\alpha,\beta) \propto y^{\alpha-1}(y-1)^{\beta-1} \tag{2.2.1}$$

with $y \in (0,1)$ and $\alpha, \beta > 0$. Both α and β are shape parameters. Unfortunately shape parameters are difficult to interpret in terms of conditional expectations and thus a reparametrization corresponding to the Generalized Linear Model convention is more practical and it can be obtained by setting $\mu = \frac{\alpha}{\alpha+\beta}$ and $\phi = \alpha + \beta$:

$$f(y|\alpha,\beta) \propto y^{\mu\phi-1}(y-1)^{(1-\mu)\phi-1} \tag{2.2.2}$$

with $E(y) = \mu$ and $Var(y) = \mu(1-\mu)\frac{1}{1+\phi}$ (dispersion depends partially on location). It is possible to model both the mean μ and the "precision" parameter ϕ with their own distinct sets of predictors but in our model only the mean depends on the covariates, while we presume ϕ constant. Let $y_1, y_2, ..., y_n$ denote independent beta variables, with mean μ_i and unknown precision parameter ϕ and let X be the matrix of covariates (continuous and/or categorical), with x_i being the i-th row vector of this matrix. There are several possible choices for the link function for the mean. For a comparison of the mean link functions see McCullagh and Nelder (1989). The most used link function and the one with an easier interpretation of the parameters is the logit-transformation:

$$ln(\frac{\mu_i}{1-\mu_i}) = x_i b. \tag{2.2.3}$$

Inverting the link function to give predicted values, the mean model may be written as

$$\mu_i = \frac{exp(x_i b)}{1 + exp(x_i b)}. \tag{2.2.4}$$

Note that the intercept has been incorporated into the coefficient vector. The log-likelihood function based on a sample of n independent observations is

$$ln(L(b, \phi)) = \sum_{i=1}^{n} ln(L_i(b, \phi)) \tag{2.2.5}$$

where

$$lnL_i(\mu_i, \phi) = ln\Gamma\phi - ln\Gamma(\mu_i\phi) - ln\Gamma[(1-\mu_i)\phi] + (\mu_i\phi - 1)lny_i + [(1-\mu_i)\phi - 1]ln(1-y_i) \tag{2.2.6}$$

The maximum likelihood estimators of b and ϕ are obtained by numerically maximizing the log-likelihood function using a non-linear optimization algorithm.

One limitation of the beta regression is that boundary values, 0 and 1, must be excluded from the dependent variable. In order to avoid obtaining distances equal to 0 we choose a template with 145 statuses instead of 144 and thus it is impossible to have

the 0 value. Following the approach suggested by Smithson and Verkuilen (2006), we avoid 1 by taking a weighting average $d'' = (d'(n-1)+s)/n$, where n is the sample size and $s \in [0,1]$ and usually it is set equal to 0.5.

Fractional logit has been proposed by Papke and Woldrige (1996) as an alternative to beta regression models for models with a fractional dependent variable. Unlike beta regression, it allows to handle data at extreme value of $[0,1]$. It consists of a quasi-likelihood estimation method for regression models y, $0 \leq y \leq 1$. As in the previous model, to explain the dependent variable y trough a vector of explanatory variables x, we assume that the conditional expected value of y given x can be written as

$$E(y_i|x_i) = G(x_i b) \tag{2.2.7}$$

where $G(\bullet)$ satisfies $0 < G(z) < 1$, $\forall z \in \mathbb{R}$. We will use a logistic function for G. The estimation procedure proposed is based on the Bernoulli quasi-likelihood method. The quasi-maximum likelihood estimator of b is obtained by maximizing the Bernoulli log-likelihood function

$$lnL(b) = \sum_{i=1}^{n} lnL_i(b) \tag{2.2.8}$$

where

$$lnL_i(b) = y_i lnG(x_i b) + (1-y_i) ln(1-G(x_i b)). \tag{2.2.9}$$

2.3 Results and discussion

With both models, we analyse the impact of the four covariates (Education, Religion, Divorce, Cohort) on the distance of each sequence in the sample to the template Single with at least one child from the age of 18 (SC/145). Tables 2.2-2.10 display the results of the estimation of country specific models, using $LLCS$, NMS, and OM [2] distances

[2]See Section 1.3 for the description of the metrics

for the fractional logit model. Results from beta regression model are similar and we omit them here for brevity. [3].

SWEDEN			
	Fractional logit		
COVARIATES	LLCS	NMS	OM
Educ 2	-0.104	0.062	**0.387**
Educ 3	0.119	0.125	**0.605**
Religion	-	-	-
Divorce	**-0.274**	0.192	**-0.418**
Cohort	0.022	**0.187**	**0.165**
Constant	**1.005**	**-2.960**	**0.407**

Table 2.2: Fractional logit model-coefficients for Sweden.
Boldface indicates p-values lower than 0.05

FINLAND			
	Fractional logit		
COVARIATES	LLCS	NMS	OM
Educ 2	0.033	0.205	**0.227**
Educ 3	0.030	-0.019	**0.752**
Religion	-	-	-
Divorce	-0.135	0.429	-0.093
Cohort	-0.012	0.169	0.010
Constant	**0.978**	**-3.120**	**0.651**

Table 2.3: Fractional logit model-coefficients for Finland.
Boldface indicates p-values lower than 0.05

[3]For a comparison of the results of the two model see Appendix

USA			
	Fractional logit		
COVARIATES	LLCS	NMS	OM
Educ 2	**0.357**	-0.310	**0.696**
Educ 3	**0.333**	-0.098	**0.839**
Religion	0.005	**-0.515**	**-0.303**
Divorce	**-0.114**	0.242	**-0.165**
Cohort	0.016	0.095	**0.080**
Constant	**0.546**	**-3.231**	**0.370**

Table 2.4: Fractional logit model-coefficients for USA.
Boldface indicates p-values lower than 0.05

CANADA			
	Fractional logit		
COVARIATES	LLCS	NMS	OM
Educ 2	-	-	-
Educ 3	-	-	-
Religion	-0.017	**0.621**	-0.076
Divorce	-0.175	0.069	**-0.233**
Cohort	-0.073	0.015	0.032
Constant	**1.106**	**-3.501**	**1.036**

Table 2.5: Fractional logit model-coefficients for Canada.
Boldface indicates p-values lower than 0.05

POLAND			
	Fractional logit		
COVARIATES	LLCS	NMS	OM
Educ 2	0.150	-0.018	-0.008
Educ 3	0.079	-0.043	**0.586**
Religion	-0.162	0.071	-0.490
Divorce	0.039	0.187	0.034
Cohort	0.022	-0.049	0.066
Constant	**1.235**	**-1.803**	0.139

Table 2.6: Fractional logit model-coefficients for Poland.
Boldface indicates p-values lower than 0.05

ESTONIA			
	Fractional logit		
COVARIATES	LLCS	NMS	OM
Educ 2	0.152	**-0.737**	0.178
Educ 3	**0.277**	**-1.083**	**0.458**
Religion	-0.092	0.263	-0.039
Divorce	**-0.365**	0.319	**-0.240**
Cohort	-0.021	0.013	-0.037
Constant	**0.777**	**-2.138**	**0.371**

Table 2.7: Fractional logit model-coefficients for Estonia.
Boldface indicates p-values lower than 0.05

The reader immediately notes that different metrics lead to different results. To understand the reason for these differences and to assess on which one is the preferable distance, we first show the graphs for the three distances for the Netherlands (similar graphs can be presented for the other countries) and discuss the causes of these differences.

From Figure 2.2 one notes that the NMS distance cannot be seen as correctly picking up the distance we are studying because we chose a template describing a very rare situation, while the NMS distance graph shows a completely different scenario in which most of the sequences seem to be very close to the template. The $LLCS$ distance

NETHERLANDS			
	Fractional logit		
COVARIATES	LLCS	NMS	OM
Educ 2	-0.077	0.114	**0.269**
Educ 3	-0.032	0.107	**0.981**
Religion	**0.076**	-0.113	-0.030
Divorce	**-0.161**	0.293	**-0.255**
Cohort	-0.062	**0.165**	**0.235**
Constant	**1.071**	**-2.960**	**0.657**

Table 2.8: Fractional logit model-coefficients for Sweden.
Boldface indicates p-values lower than 0.05

ITALY			
	Fractional logit		
COVARIATES	LLCS	NMS	OM
Educ 2	**-0.143**	-0.030	**0.657**
Educ 3	**0.152**	**-0.111**	**1.228**
Religion	0.122	-0.069	**-0.399**
Divorce	-0.256	0.268	-0.112
Cohort	**0.122**	-0.045	**0.177**
Constant	**1.058**	**-1.921**	**-0.190**

Table 2.9: Fractional logit model-coefficients for Italy.
Boldface indicates p-values lower than 0.05

produces a more coherent graphic; and the most coherent is from the OM distance. From some descriptive analyses (not shown) we verified that the NMS distance is very sensitive to the length of the sequence: distance between very long sequences and the template are extremely big. So that the transformation $d' = (d - m)/(M - m) \approx d/M$ rescales most distances close to zero. $LLCS$ yields more homogeneous values. Moreover, SC is a status visited by very few individuals and this can be a limit to the way in which both NMS and $LLCS$ are obtained, bearing in mind that SC is the only subsequence that the template can have in common with any other sequence. Another reason why results of the NMS and $LLCS$ distances are not coherent is because all

SPAIN			
	Fractional logit		
COVARIATES	LLCS	NMS	OM
Educ 2	0.043	-0.041	**0.260**
Educ 3	0.057	-0.008	**0.686**
Religion	0.141	**0.022**	-0.046
Divorce	-0.228	0.182	-0.296
Cohort	0.080	0.049	0.023
Constant	**0.957**	**-2.476**	**0.538**

Table 2.10: Fractional logit model-coefficients for Spain.
Boldface indicates p-values lower than 0.05

statuses other than SC are all treated in the same manner. We believe that for example, being in unmarried cohabitation with a child is closer to the status of single mother with a child, than to the status of married without children and a distance should consider this aspect. Looking at Figure 2.3, that shows the sequences ordered according to the distance from the template SC/145, we note that only with OM we have a rational order of sequences: the sequences with a "C" state and with a longer permanence in this state are closer to SC/145. The NMS and $LLCS$ distances are probably not suitable in a template with only one status which occurs very rarely in our sample. We believe that a template with more status options would generate more meaningful values.

The distribution of the distance from SC/145 obtained with OM is coherent with what we expected. Therefore, in order to understand how different covariates effect the risk of premarital motherhood we concentrate our attention on OM results (column 4 of each table).

As already noted in literature, the estimates suggest that education influences behavior that affects single childbearing. All countries have statistically significant positive "Educ3"-coefficients. This suggests that longer education, after the age of 18, compared to no education after the age of 15, tends to increase the motivation to control

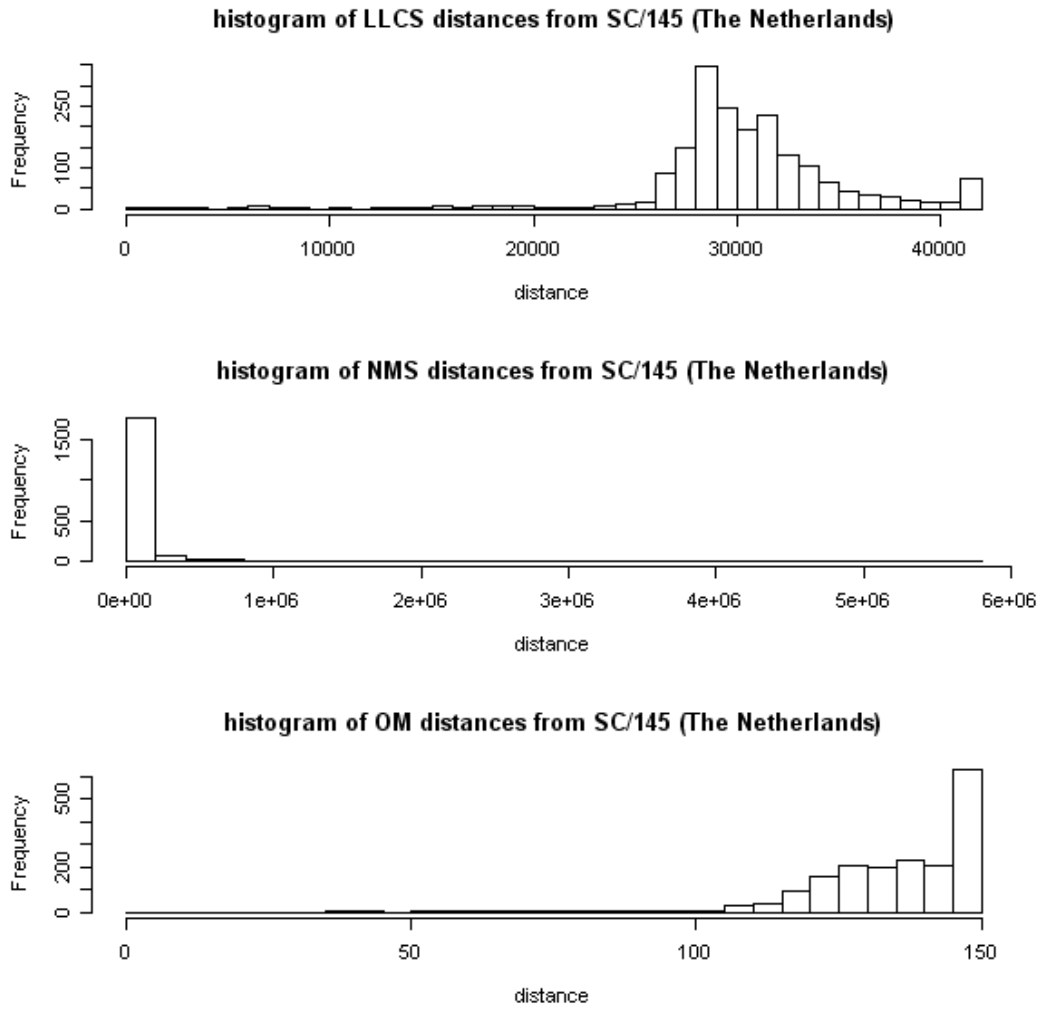

Figure 2.2: Histograms of the distance from SC/145 template, computed with different metrics for the Netherlands.

fertility. "Educ2" has an impact in the same direction, but smaller, in all countries, with the exception of Poland and Estonia where the coefficient is not significant. Religiousness is negatively associated with the distance from the template SC/145 thus this covariate seems to have the effect of increasing the propensity of teenage childbearing. The "Religion" coefficient has the same negative sign for each country, although it is statistically significant only for Italy, the Netherlands and USA. If we look at the magnitude of this factor in the different countries we note that only in Poland, Italy,

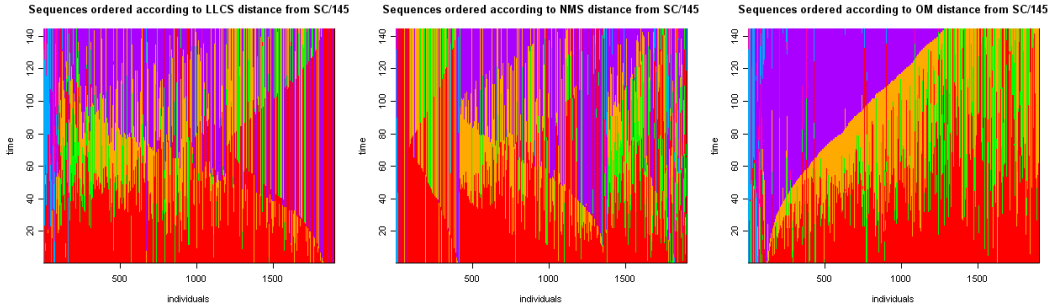

Figure 2.3: Representation of individual sequences ordered according to the distance from the template SC/145. Time in months is on the vertical axis and the number of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC).

and USA, in which traditionally the influence of the Catholic Church is strong, religiosity has a substantial impact.

The family background characteristic has the expected influence on the distance from the template SC/145. Results suggest that having divorced or separated parents is a significant determinant in decreasing this distance. The "Divorce" coefficient is negative for all countries, with the only exception of Poland, where probably the number of "divorced cases" is insufficient for a correct estimate of the parameter. The impact of divorced or separated parents has a stronger and statistically significant effect in Sweden, Estonia, Canada, and USA.

The distance from the single mother template is significantly related also to "Cohort". The younger cohort seems to be less subject to the propensity of being a single mother. Only in Estonia and Finland is there an effect in the opposite direction, but the coefficient is not significant and very close to zero. In all other countries the coefficient is positive, but significant only in Italy, the Netherlands, Sweden and USA. Compared to other factors, "Cohort" has the smallest influence on the distance to the SC/145 template, while education has the strongest impact.

2.4 Concluding remarks

The focus of this chapter has been to investigate the effects of several individual and family background characteristics on the distance from the teenage and long run single mother pattern in nine European and North American countries. In order to do this, we defined a template trajectory that describes the "risky mother", calculated the distance between individual sequences and this trajectory. This distance was used as a dependent variable in beta regression and in a fractional logit model. The evidence suggests that education, religiosity, familiar instability, and cohort are significantly related to teenage premarital pregnancy. In particular, long run education enrollment and cohort are negatively associated with the propensity to being a long run single mother. This is not unexpected since younger cohort of women shows higher percentages participating in education and prolonged education period. On the contrary, religiosity and parental divorce are positively related to this propensity. Education has a significant effect in all countries and the strongest influence. From a policy perspective, interventions to encourage pursuit of higher levels of education could have a role in reducing the prevalence of teenage childbearing.

Looking at the level of magnitude and at the significant factors (but not at the sign), we find heterogeneous results in different countries. National cultural aspects seem to have an influence in sexual, contraceptive, and abortion attitudes.

Our analysis also shows that different metrics lead to relevant differences in the results. Often the importance of the choice of a metric is underestimated and metric is considered as given. On the contrary, there is no external criterion to evaluate them and thus it is impossible to establish *a priori* which metric is preferable with respect to the others. Preliminary analysis about this choice is crucial to obtain meaningful results. We show that the distant distribution obtained with OM is coherent with our

data, while NMS and $LLCS$ attributes are less suitable. The principal limitation of NMS metric is that distances between very long sequences and the template are extremely big. Unfortunately, to prevent this problem it is not possible using normalized distances because normalization leads to have so small values that the precision of the resulting decimal representation is not sufficient and most of the values result equal to zero. Both the metrics are affected by the fact that we have a template with only one state, bearing in mind that SC is the only subsequence that the template can have in common with any other sequence. A possible solution to these problems could be to choose a template with several states but this choice would not be coherent with the object of the analysis: young single mothers who remain permanently in this condition. Moreover, both the metrics consider all statuses other than SC in the same manner, but in a study like this it is reasonable to make differences among the states, e.g. to consider SC closer to UC than to M. A more appropriate and useful attribute could be:

$$A(s, s') = \sum_{i=1}^{n} w(s_i, s'_i) \tag{2.4.1}$$

with $w(s_i, s'_i) = 1$ if $s_i = s'_i$ and $0 \leq w(s_i, s'_i) < 1$ if $s_i \neq s'_i$. It easy to show that if the coefficients $w(s_i, s'_i)$ are symmetric, $d(s, s') = A(s, s) + A(s', s') - 2A(s, s')$ is a metric, of which the Hamming distance is a special case.

These considerations, although very important, go beyond the scope of this chapter, in which we wanted to propose regression methods to detect subgroups as potential targets of specific policies and to draw the attention on the importance of the metric's choice.

2.5 Appendix

The results obtained with the different regression models agree and they lead to the same credible conclusions about which factors contribute to explain the distance from the adolescent single mother pattern.

SWEDEN						
	Fractional logit			Beta regression		
COVARIATES	LLCS	NMS	OM	LLCS	NMS	OM
Educ 2	-0.104	0.062	**0.387**	-0.161	0.037	**0.314**
Educ 3	0.119	0.125	**0.605**	-0.160	0.032	**0.487**
Religion	-	-	-	-	-	-
Divorce	**-0.274**	0.192	**-0.418**	**-0.318**	0.014	**-0.280**
Cohort	0.022	**0.187**	**0.165**	-0.014	0.046	**0.136**
Constant	**1.005**	**-2.960**	**0.407**	**1.130**	**-2.541**	**0.531**

Table 2.11: Fractional logit and beta regression model-coefficients for Sweden.
Boldface indicates p-values lower than 0.05

FINLAND						
	Fractional logit			Beta regression		
COVARIATES	LLCS	NMS	OM	LLCS	NMS	OM
Educ 2	0.033	0.205	**0.227**	0.069	0.070	**0.233**
Educ 3	0.030	-0.019	**0.752**	0.070	-0.013	**0.614**
Religion	-	-	-	-	-	-
Divorce	-0.135	0.429	-0.093	0.097	0.092	0.064
Cohort	-0.012	0.169	0.010	-0.021	0.052	0.027
Constant	**0.978**	**-3.120**	**0.651**	**1.008**	**-2.762**	**0.746**

Table 2.12: Fractional logit and beta regression model-coefficients for Finland.
Boldface indicates p-values lower than 0.05

USA						
	Fractional logit			Beta regression		
COVARIATES	LLCS	NMS	OM	LLCS	NMS	OM
Educ 2	**0.357**	-0.310	**0.696**	-0.056	0.105	0.437
Educ 3	**0.333**	-0.098	**0.839**	**0.339**	-0.050	**0.625**
Religion	0.005	**-0.515**	**-0.303**	0.012	**-0.141**	**-0.244**
Divorce	**-0.114**	0.242	**-0.165**	**-0.128**	**0.107**	**-0.084**
Cohort	0.016	0.095	**0.080**	0.064	0.020	0.036
Constant	**0.546**	**-3.231**	**0.370**	**0.438**	**-3.304**	**0.489**

Table 2.13: Fractional logit and beta regression model-coefficients for USA.
Boldface indicates p-values lower than 0.05

CANADA						
	Fractional logit			Beta regression		
COVARIATES	LLCS	NMS	OM	LLCS	NMS	OM
Educ 2	-	-	-	-	-	-
Educ 3	-	-	-	-	-	-
Religion	-0.017	**0.621**	-0.076	**-0.394**	**0.320**	-0.244
Divorce	-0.175	0.069	**-0.233**	-0.049	0.027	-0.077
Cohort	-0.073	0.015	0.032	-0.030	0.004	0.041
Constant	**1.106**	**-3.501**	**1.036**	**1.436**	**-3.110**	**0.925**

Table 2.14: Fractional logit and beta regression model-coefficients for Canada.
Boldface indicates p-values lower than 0.05

POLAND						
	Fractional logit			Beta regression		
COVARIATES	LLCS	NMS	OM	LLCS	NMS	OM
Educ 2	0.150	-0.018	-0.008	0.090	-0.035	0.019
Educ 3	0.079	-0.043	**0.586**	0.029	0.056	**0.542**
Religion	-0.162	0.071	-0.490	-0.183	0.039	-0.399
Divorce	0.039	0.187	0.034	-0.194	0.166	0.091
Cohort	0.022	-0.049	0.066	0.058	-0.058	0.070
Constant	**1.235**	**-1.803**	0.139	**1.161**	**-1.794**	0.141

Table 2.15: Fractional logit and beta regression model-coefficients for Poland.
Boldface indicates p-values lower than 0.05

ESTONIA						
	Fractional logit			Beta regression		
COVARIATES	LLCS	NMS	OM	LLCS	NMS	OM
Educ 2	0.152	**-0.737**	0.178	0.184	**-0.384**	0.149
Educ 3	**0.277**	**-1.083**	**0.458**	**0.338**	**-0.437**	**0.393**
Religion	-0.092	0.263	-0.039	-0.062	0.055	-0.002
Divorce	**-0.365**	0.319	**-0.240**	**-0.407**	0.146	**-0.246**
Cohort	-0.021	0.013	-0.037	-0.041	-0.015	-0.044
Constant	**0.777**	**-2.138**	**0.371**	**0.747**	**2.060**	**0.442**

Table 2.16: Fractional logit and beta regression model-coefficients for Estonia.
Boldface indicates p-values lower than 0.05

NETHERLANDS						
	Fractional logit			Beta regression		
COVARIATES	LLCS	NMS	OM	LLCS	NMS	OM
Educ 2	-0.077	0.114	**0.269**	-0.070	0.045	**0.231**
Educ 3	-0.032	0.107	**0.981**	0.065	0.012	**0.693**
Religion	**0.076**	-0.113	-0.030	0.067	-0.065	**-0.101**
Divorce	**-0.161**	0.293	**-0.255**	-0.078	**0.160**	**0.118**
Cohort	-0.062	**0.165**	**0.235**	**-0.162**	**0.118**	**0.168**
Constant	**1.071**	**-2.960**	**0.657**	**1.130**	**-2.541**	**0.934**

Table 2.17: Fractional logit and beta regression model-coefficients for Sweden.
Boldface indicates p-values lower than 0.05

ITALY						
	Fractional logit			Beta regression		
COVARIATES	LLCS	NMS	OM	LLCS	NMS	OM
Educ 2	**-0.143**	-0.030	**0.657**	-0.059	-0.035	**0.576**
Educ 3	**0.152**	**-0.111**	**1.228**	**0.335**	**0.116**	**0.998**
Religion	0.122	-0.069	**-0.399**	0.139	0.068	**-0.271**
Divorce	-0.256	0.268	-0.112	0.192	-0.152	-0.112
Cohort	**0.122**	-0.045	**0.177**	**0.172**	0.048	**0.177**
Constant	**1.058**	**-1.921**	**-0.190**	**0.940**	**-1.866**	0.169

Table 2.18: Fractional logit and beta regression model-coefficients for Italy.
Boldface indicates p-values lower than 0.05

SPAIN						
	Fractional logit			Beta regression		
COVARIATES	LCS	NMS	OM	LCS	NMS	OM
Educ 2	0.043	-0.041	**0.260**	0.163	-0.033	**0.295**
Educ 3	0.057	-0.008	**0.686**	**0.274**	-0.044	**0.601**
Religion	0.141	**0.022**	-0.046	0.021	0.058	-0.105
Divorce	-0.228	0.182	-0.296	-0.148	0.033	-0.184
Cohort	0.080	0.049	0.023	0.059	0.028	0.049
Constant	**0.957**	**-2.476**	**0.538**	**1.041**	**-2.397**	**0.632**

Table 2.19: Fractional logit and beta regression model-coefficients for Spain.
Boldface indicates p-values lower than 0.05

Chapter 3

Analysis of dispersion (ANODI) with permutation test

3.1 Introduction

The description of life courses has attracted increasing interest in the literature. There are a variety of methodological approaches to the analysis of sequences. Most attention in literature has been devoted to finding and efficiently describing common patterns among sequences. However, few results were obtained to interpret underlying factors driving demographic behaviour. One of the principal contributions is that of McVicar and Anyadike-Danes (2002), who suggested to use clusters obtained with Optimal Matching as the dependent variable in a multinomial logit model. The aim is to explain cluster membership as a function of a set of socio-economic variables. One of the drawbacks of that method is the excessive information loss in the clusters' membership: only a fraction of the relationship between the explicative structure and the transformed simplified data can be explained. Moreover, clusters contain only a fraction of the information present in the full sequence data. We propose a method that, following the same logic of the multivariate Analysis of variance (ANOVA), allows the evaluation of the effect of one factor on sequence distance. The criterion may also be used to build a parsimonious model explaining the relationship between the

sequences and the explanatory variables. The advantage compared to a multinomial model applied to clusters is that we infer the relationship between the sequences and the explanatory structure not from possibly overly simplified data, i.e. clusters, but from the distance matrix which contains more information on the original sequences structure.

The proposed technique is applied to the Dutch Fertility and Family Surveys (FFS) dataset. The distances between sequences are calculated using the NMS metric proposed by Elzinga (2003, 2005), and described in Chapter 1. This metric depends on the number of common subsequences of a pair of sequences, weighted by the frequency of the occurrence of these subsequences in both sequences. Of course, these distances cannot be seen as Euclidean distances calculated on the basis of iid normally distributed variables, as in the traditional ANOVA setting.

To evaluate the relationship between distances and the socio-demographic variables, we adapt the (multivariate) ANOVA approach to our sequence context or, better, to the situation when the unique information about the response variable of interest is synthesized in a dissimilarity matrix. Permutation tests are used to evaluate the significance of an effect. Besides this "marginal" analysis, i.e., besides the evaluation of the significance of each factor on sequences, we also evaluate partial effects, i.e., the significance of one effect given that other effects are taken into account. On the basis of these measures we can introduce a model building procedure, aiming at selecting the most relevant factors describing the observed life courses.

The chapter is organized as follows. In Section 2 we introduce and describe the Analysis of Dispersion. The adequacy of the method is evaluated via simulation studies in Section 3. Section 4 is devoted to the presentation of the selection model procedure. In Section 5 we apply the proposed method to FFS data. Section 6 summarizes the main findings.

3.2 Analysis of dispersion and permutation test

Analysis of variance aims at evaluating the impact of a qualitative variable X on a normally distributed variable Y. As mentioned above, the traditional ANOVA approach can not be used to analyze our FFS data. Nevertheless, as it will be illustrated later, the quantities at the basis of the classical ANOVA approach can be rewritten in terms of distances between couples of cases. Hence, ANOVA can be easily extended to the context under analysis. To evaluate the significance of the statistics, i.e. to test the hypothesis of null effect, we refer to permutation tests. With a permutation test we obtain the reference distribution of the test statistic in question first by rearranging the observations of the variable we are analysing, then by calculating the value of this test statistic for the rearranged (permutated) sample, and repeating this procedure many times (e.g. 5000 times). For a detailed review of permutation tests and of their property, see Pesarin (2001). The procedure can be described as follows:

1. choose a statistic that reflects ANOVA logic, in terms of dispersion and not variance, to measure the effect of one factor on the dispersion;

2. obtain the sampling distribution of the statistic under the null hypothesis of no effect;

3. evaluate the p-value of the observed statistic under the null hypothesis.

For the first step, we exploit the fact that in ANOVA there exists a correspondence between variance and distances, i.e. the variation within groups is proportional to the Euclidean distance within groups.

Before proceeding further, we briefly recall the main ANOVA concepts which are strictly connected to the statistic we will introduce (see e.g. Jobson, 1992). Given the g groups induced by one factor X, we denote by y_{ij} the i-th observation on Y taken in group

j where $i = 1, 2, ..., n_j$ and $j = 1, 2, ..., g$. The total variation in Y over the sample, (TSS or the total sum of squares), is described by $TSS = \sum_{j=1}^{g} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{..})^2$. The sample means for the g groups are denoted by $\bar{y}_{.1}, \bar{y}_{.2}, ..., \bar{y}_{.g}$ where $\bar{y}_{.j} = \sum_{i=1}^{n_j} y_{ij}/n_j$. The mean of all n observations is the grand mean $\bar{y}_{..} = \sum_{j=1}^{g} \sum_{i=1}^{n_j} y_{ij}/n$. The total variation within the g groups (the within groups sum of squares) is given by $WSS = \sum_{j=1}^{g} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{.j})^2$, while the variation explained by the fitted model (the between groups sum of squares) is given by $BSS = \sum_{j=1}^{g} \sum_{i=1}^{n_j} (\bar{y}_{.j} - \bar{y}_{..})^2$. The variance due to the differences within individual samples is the within group variation divided by its degrees of freedom, $WSS/(n-g)$ and the variance due to the interaction between the samples is the between group variation divided by its degrees of freedom, $BSS/(g-1)$. The F test statistic is defined as $F = \frac{BSS/(g-1)}{WSS/(n-g)}$.

It is known that WSS is proportional to the Euclidean squared distance among observations within all groups: $WSS \propto \sum_{j=1}^{g} \sum_{i=1}^{n_j} \sum_{h=1}^{n_j} (y_{ij} - y_{hj})^2$ and on this property we base our test. Following the definitions of the total dispersion and the dispersion within groups proposed in Piccarreta and Billari (2007), we set $W_D = \sum_{j=1}^{g} \sum_{i=1}^{n_j} \sum_{h=1}^{n_j} (d_{NMS}(i,h))^2$, $T_D = \sum_{i=1}^{n} \sum_{h=1}^{n} (d_{NMS}(i,h))^2$ and $B_D = T_D - W_D$. Therefore we utilize the ratio $T = B_D/W_D$ as the test statistic of the Analysis of Dispersion. As concerns the calculation of the p-value, we notice that $T = B_D/W_D$ is equivalent to $\frac{B_D/(g-1)}{W_d/(n-g)}$, since as $(n-g)/(g-1)$ is a constant multiplier over all the data permutations with respect to the test statistic value and therefore it has no effect on the p-value.

In general, this test is realized by a simulation of the testing problem conditional on the observed distance data set D, that is, by a without-replacement resampling procedure. Essentially the method operates according to the following steps:

1. calculate for the given data set D, the test statistic $T : T_0 = T(D)$;

2. consider a random attribution of the order of the observations, obtaining a permuted data set D^*;

3. calculate $T_s = T(D^*)$;

4. repeat steps 2 and 3 for a total of K times.

The K permutation sets D^* are a random sample from the permutation sample space. Thus, the K corresponding values of T_s simulate the null permutation distribution of T. Therefore, they permit the statistical estimation of the permutation c.d.f. $F(d|D)$ and the significance level function $L(z|D) = Pr(T_0 \geq z|D)$ respectively by the empirical distribution function $\hat{F}_K(z) = \#(T_s \leq z)/K$ and by $\hat{L}^*_K(z) = \#(T_s \geq z)/K$, $\forall z \in \mathbb{R}$. In practice, the estimated p-value corresponding to the significance level evaluated on the observed value T_0 is given by $\lambda = \hat{L}^*_K(T_0) = \#(T_s \geq T_0)/K$. If $\lambda \leq \alpha$, we may conclude that the data disagree with H_0.

This procedure, although employing only a sample of the possible data permutations, is appropriate because it respects the validity criterion for randomization tests (Edgington, 2007): "a statistical testing procedure is valid if, under the null hypothesis, the probability of a p-value as small as p is no greater than p, for any p". However, if H_0 is false and there is an actual factor effect, it could be less powerful than using all the possible data permutations because of an increased discreteness in the reference set and a larger minimum p-value. Increasing the number of data permutations makes the difference in power less influential.

3.3 Simulation study

We perform a limited simulation study to assess the efficacy of the method under a variety of conditions. The data is generated by a model similar to that implemented in Chapter 4. In particular, in the simulated data we examine:

1. how our method captures the presence of structure;

2. how the performance varies with the number of the groups g of X;

3. how the performances varies with the probability of belonging to each group.

For each trial in each simulation, we generate 1000 sequences, which reproduce sequences of length similar to those of FFS data, with a parametric model in which the transition probabilities and the duration distribution between subsequent transitions from state to state depend on a covariate through the logit model. They represent the sequences with structure. We compare the results obtained using a sample of 1000 sequences with structure with those from a sample of 1000 sequences without structure, i.e. in which the distance between sequences has a random relationship with the covariate. For each simulation 1000 trials are performed.

We assume that $\{S_k, k = 1, ..., r+1\}$ indicate the states that the individual experiences, in the order in which they have been visited, and $\{T_k, k = 1, ..., r+1\}$ be the corresponding times spent in the $r+1$ states. Let X denote the matrix of the covariate. The dimension of the matrix depends on the number of groups g. Using the first category as the baseline, the other $g-1$ groups are converted into $g-1$ dummy variables. Since we incorporate the intercept in the model the number of columns is equal to g.

We assume that T_k follows a geometric distribution with parameter p that depends on the covariate X through the logit link:

$$p(X) = \frac{exp(\beta' X)}{1 + exp(\beta' X)} \tag{3.3.1}$$

We define the transition probabilities between from state i_k to state j_k at the k-th transition, depending on covariates, by:

$$P_{ijk} = Pr(S_k = j | S_{k-1} = i, X) \tag{3.3.2}$$

where $i, j = \{1, ..., J\}$ and $k = 1, ..., K$, with K representing the maximum number of possible transitions, J the total number of states, and $P_{iik} = 0$, $\forall i$, $\forall k$. The transition probabilities can be modeled through generalized logits. Using the g-th category as a reference, the logits can be parametrized as

$$P_{ijk} = \frac{exp(\gamma'_{ij}X)}{1 + \sum_{h=\{1,...,g-1\}-\{i\}} exp(\gamma'_{ih}X)} \tag{3.3.3}$$

with $j = \{1, ..., g-1\} - \{i\}$ and

$$P_{igk} = \frac{1}{1 + \sum_{h=\{1,...,g-1\}-\{i\}} (exp(\gamma'_{ih}X))}. \tag{3.3.4}$$

To generate the covariate we utilize a discrete distribution

$$p(X = k) = \begin{cases} 1 & prob = p_1 \\ . & . \\ . & . \\ . & . \\ g & prob = p_g \end{cases} \qquad k = 1, ..., g \tag{3.3.5}$$

It is a simplified (the covariate is not time-varying) version of the model presented in Chapter 4. For all simulations, in the transition submodel we keep the intercept coefficients constant and equal to 1 and the transition coefficients equal to -2 for all possible transitions and dummy-covariates, i.e. $\gamma_{0ij} = 1$ $\gamma_{lij} = -2$, with $i = 1, ..., 6$, $j = 1, ..., 4$ and $l = 1, ..., g-1$. Taking into account that we have six distinct states and the sixth is the reference state, the transition matrices of an individual with $X = 0$

and $X = 1$ are respectively:

$$P_{ij} = \left\{ \begin{array}{cccccc} State & I & II & III & IV & V & VI \\ I & 0.0 & 0.23 & 0.23 & 0.23 & 0.23 & 0.08 \\ II & 0.23 & 0.0 & 0.23 & 0.23 & 0.23 & 0.08 \\ III & 0.23 & 0.23 & 0.0 & 0.23 & 0.23 & 0.08 \\ IV & 0.23 & 0.23 & 0.23 & 0.0 & 0.23 & 0.08 \\ V & 0.23 & 0.23 & 0.23 & 0.23 & 0.0 & 0.08 \\ VI & 0.23 & 0.23 & 0.23 & 0.23 & 0.08 & 0.0 \end{array} \right\} \tag{3.3.6}$$

$$P_{ij} = \left\{ \begin{array}{cccccc} State & I & II & III & IV & V & VI \\ I & 0.0 & 0.15 & 0.15 & 0.15 & 0.15 & 0.4 \\ II & 0.15 & 0.0 & 0.15 & 0.15 & 0.15 & 0.4 \\ III & 0.15 & 0.15 & 0.0 & 0.15 & 0.15 & 0.4 \\ IV & 0.15 & 0.15 & 0.15 & 0.0 & 0.15 & 0.4 \\ V & 0.15 & 0.15 & 0.15 & 0.15 & 0.0 & 0.4 \\ VI & 0.15 & 0.15 & 0.15 & 0.15 & 0.4 & 0.0 \end{array} \right\} \tag{3.3.7}$$

Note that the rows of the transition matrix are conditional probabilities that must add to 1. In the duration submodel, we set the coefficients equal to $\beta = (-4, 1)$, $\beta = (-4, 0.5, 1)$, $\beta = (-4, 0.4, 0.7, 0.9)$, $\beta = (-4, 0.4, 0.7, 0.9, 1.1)$, respectively for the models in which the covariate groups are 2, 3, 4, 5. Starting from the distance matrix, D_{NMS}, we calculate the value of the test statistic $T_0 = B_D/W_D$. We achieve a permutation sample, by assigning at random each individual to a value of the covariate, while keeping the dissimilarity matrix unchanged. In this way, we are assigning X at random and consequently we are considering the situation in which X has no effect on sequences and thus on dissimilarities. Then, we compute the statistical test T on the permuted sample. At this stage, we repeat the permutation 2000 times, and thus estimate the permutation distribution of T. The p-value is the proportion of those 2000 data permutations with a value of T larger than the value for the experimental result

T_0.

To verify that ANODI correctly captures the structure in the data, we compare the results obtained from a sample of sequences with structure with the results obtained from a sample of sequences in which the sequences structure has no relationship with the covariate.

Let X be a binary covariate, generated from a Bernoulli distribution with parameter 0.4. For each trial, we generate a sample of 1000 sequences with the logit-model approach presented above, we obtain the distance matrix, D, and we compute the statistic value $T_0 = T(D)$. We obtain the null permutation distribution of T using the values of T_s calculated on 2000 permutated data sets of the distance matrix. The p-value is the proportion of permutations with a value of T larger than T_0.

We then consider a permutation sample D^*, in which by definition there is no structure, as the distance matrix of sequences in the case of no relationship with the covariate and obtain the relative p-value as the proportion of permutated samples with a value of T larger than $T(D^*)$. We replicate this whole procedure for 1000 times.

We can now compare the p-values obtained from the sequences with structure with the p-values obtained from the sequences without structure. Figure 3.1 shows the histograms of the obtained p-values. Our results provide interesting evidence of the effectiveness of the method: all the trials from structured sequences report a p-value equal to 0 (reject H_0), while from unstructured data we register different values in the interval [0,1], consistent with the absence of relationship (recall that the p-value follows a Uniform [0,1] distribution under H_0).

In the second part of the simulations, we explore how the number of groups of the covariate influences the performance of the method. We generated a covariate with

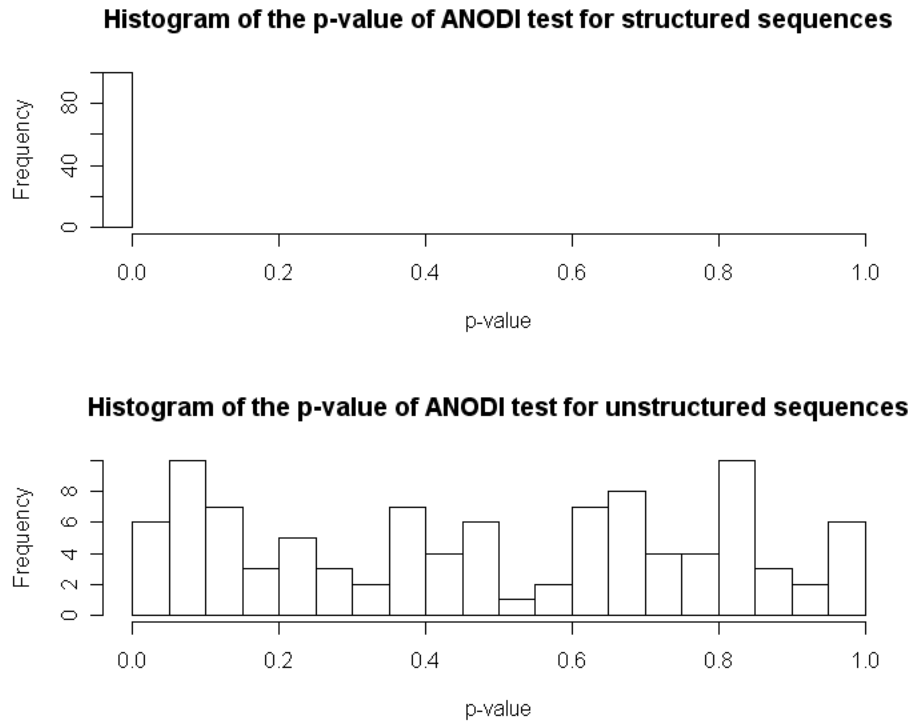

Figure 3.1: Histograms of the p-values obtained with ANODI test with "structured" and "unstructured" sequences, for 1000 trials based on 2000 permutated samples.

three, four and five groups from the following discrete distributions, respectively:

$$X_3 = \begin{cases} 1 & prob = 0.4 \\ 2 & prob = 0.3 \\ 3 & prob = 0.3 \end{cases} \tag{3.3.8}$$

$$X_4 = \begin{cases} 1 & prob = 0.4 \\ 2 & prob = 0.2 \\ 3 & prob = 0.2 \\ 4 & prob = 0.2 \end{cases} \tag{3.3.9}$$

$$X_5 = \begin{cases} 1 & prob = 0.4 \\ 2 & prob = 0.15 \\ 3 & prob = 0.15 \\ 4 & prob = 0.15 \\ 5 & prob = 0.15 \end{cases} \tag{3.3.10}$$

For all covariates, group 1 is taken as the reference level and the other groups are indicated by dummy variables. As in the previous simulation, we compare the histogram of the p-values obtained in 1000 trials of the ANODI test on the set of 1000 sequences we generated in a manner that they have a significant relationship with the covariate with the histogram of the p-values of the set of 1000 sequences with a chance relationship with the covariate. Figure 3.2 plots the histograms for the covariate with 3, 4 and 5 groups. As can be observed, when the number of groups vary from 3 to 5, the test achieves the expected results: the p-values from structured data are all equal to 0 (reject H_0) for all the trials, while the p-values with unstructured data are rather uniformly distributed in the interval [0, 1]. This means that the test is able to recognize correctly the presence of structure in the data and the variation of the number of groups does not affect the performance of the test.

In the third part of the simulation studies we evaluate the effect of the size of the two groups in a binary covariate, varying the probability of belonging to each group. To this extent we generate three Bernoulli variables with parameter 0.5, 0.25 and 0.05. As above, we compare the results from structured and unstructured data, which are shown in Figure 3.3.

A complete simulation study of the power of approach being proposed is beyond the scope of this work. However the results suggest that the ANODI test is able to determine whether a given factor has a significant effect on the distance between sequences, also with very unbalanced group sizes.

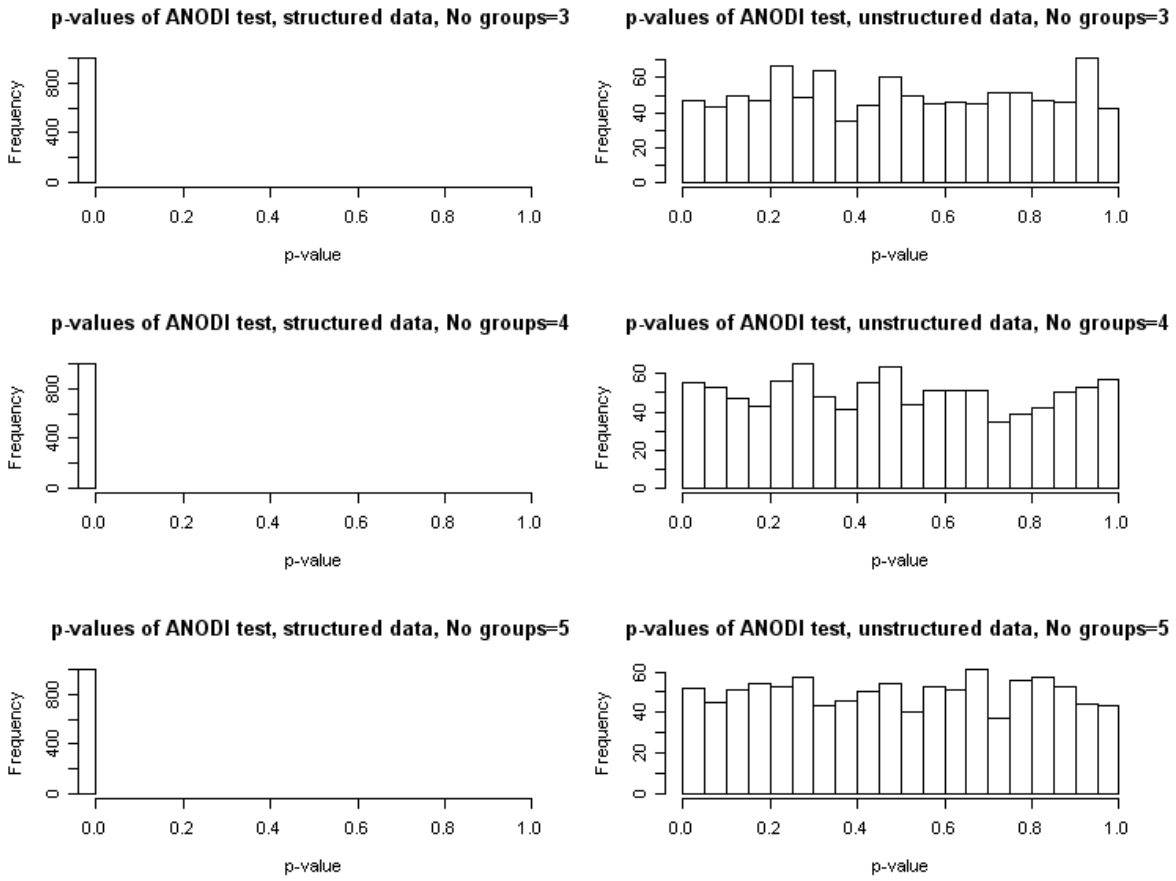

Figure 3.2: Histograms of the p-values obtained in 1000 trials with ANODI test based on 2000 permutated samples from "structured" and "unstructured" sequences when the number of groups of the covariate is 3, 4, and 5.

3.4 Model selection and permutation test

When several explanatory variables are available, the set of possible models becomes large, giving rise to a model selection problem. The selection process of the best model takes into account the usual two competing goals: the model should be complex enough to fit the data well but it should also be simple to interpret, smoothing rather than overfitting the data. In the ANOVA setting, F tests are used to compare two nested

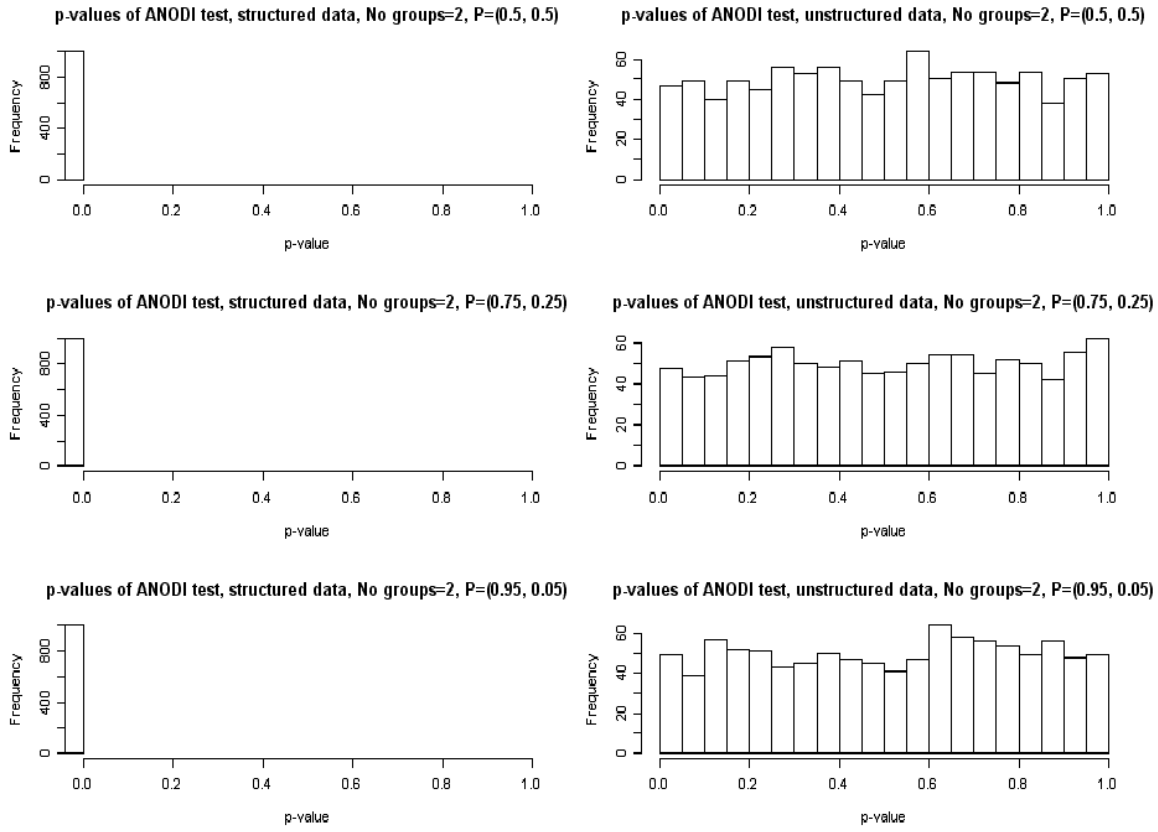

Figure 3.3: Histograms of the p-values obtained in 1000 trials with ANODI test based on 2000 permutated samples from "structured" and "unstructured" sequences for different probabilities of belonging to the two groups of the covariate: P=(0.5, 0.5), P=(0.75, 0.25) and P=(0.95, 0.05).

models with k and $k-1$ explanatory variables, with the statistic F given by:

$$F = \frac{BSS_k - BSS_{k-1}}{WSS_k/(n-k)}, \tag{3.4.1}$$

where BSS_k and BSS_{k-1} indicate the between groups sum of squares with k and $k-1$ regressors in the model, and WSS_k the within groups sum of squares with k regressors. Following the sum of squares idea we provide a statistic F^* which evaluates the reduction in the explanatory capability of the model when a certain variable j is removed. As in the usual backward elimination procedure, we start with a full model and eliminate the variables one by one from the model. At each step, the variable with the smallest

contribution to the model is tested and if its contribution results not significant it is deleted. [1]

As described in Section 2, we use a permutation procedure which operates according to the following steps:

1. calculate the observed value $F_0^* = (B_{Dk} - B_{Dk-j})/W_{Dk}$ on the given data set D;
2. consider a random attribution of the order of the observation of the variable j, obtaining a permuted data set D^* and new values for B_{Dk} and W_{Dk};
3. calculate $F^* = F^*(D^*)$;
4. repeat steps 2 and 3 K times (K=2000).

We reject the hypothesis H_0 (the removed regressor is not significant) if the proportion of F^* permutation statistics larger than F_0^* is lower than α.

By comparing the nested models, we can measure to which extent removing the tested independent variable reduces the predictability of the dependent variable. Thus our statistic is analogous to R^2 for the linear model.

3.5 ANODI of the FFS data

We now apply ANODI to Dutch Fertility and Family Survey (FFS) data. The retrospective histories of 1897 women between the age of 18 to 30 about childbearing and union formation were collected on a monthly time scale. We use some additional informations on some individual and family characteristics: level of education, religion belief, parental divorce and birth cohort. [2]The sample sizes for each category are shown

[1]It would be quite easy to obtain a forward selection procedure, but following Agresti (2002) we believe that it is "safer to delete terms from an overly complex model than to add terms to an overly simple one. Forward selection can stop prematurely because a particularly test in the sequence has low power".

[2]For the description of the data and FFS see Section 1.2

Covariates	No. categories	first category size	second category size	third category size
Educ	3	$n_1 = 181(10\%)$	$n_2 = 706(37\%)$	$n_3 = 1010(53\%)$
Religion	2	$n_0 = 707(37\%)$	$n_1 = 1190(63\%)$	-
Divorce	2	$n_0 = 1758(93\%)$	$n_1 = 139(7\%)$	-
Cohort	2	$n_1 = 915(48\%)$	$n_2 = 982(52\%)$	-

Table 3.1: Sample size of the covariates for Dutch FFS data.

Covariate	p**-value**	T_0	**(min(T^*),max(T^*))**
Educ	0	1.365	(1.295,1.352)
Religion	0	0.967	(0.863,0.899)
Divorce	0	0.183	(0.143,0.169)
Cohort	0	1.024	(0.997,1.006)

Table 3.2: Summary table for ANODI-permutation test on Dutch FFS data.
(T_0: the value of the statistic T for the original distance matrix
(min(T^*),max(T^*)): the minimum and the maximum values of the statistic T for the permutated distance matrices, T^*).

in Table 3.1.

We apply ANODI to each covariate in order to examine how the distances between cases and hence, the sequences are related to the level of education, religiousness, parental divorce and cohort. The p-values for each covariate are shown in Table 3.2. The table contains also the value of the statistic T for the observed distance matrix, T_0, and the minimum and the maximum values of the statistic for the permutated distance matrices, T_s. Figures 3.4-3.7 show the permutation distribution of the statistic T for each covariate based on 5000 permuted samples. The dashed vertical line marks the observed value T_0. Its location beyond the right tail shows that such a value is very unlikely to occur when the null hypothesis is true and thus that the effect of the covariate is significant.

Evidence suggests that all the covariates we consider affect the process of family formation in the Netherlands. We can graphically inspect the effect of the covariates,

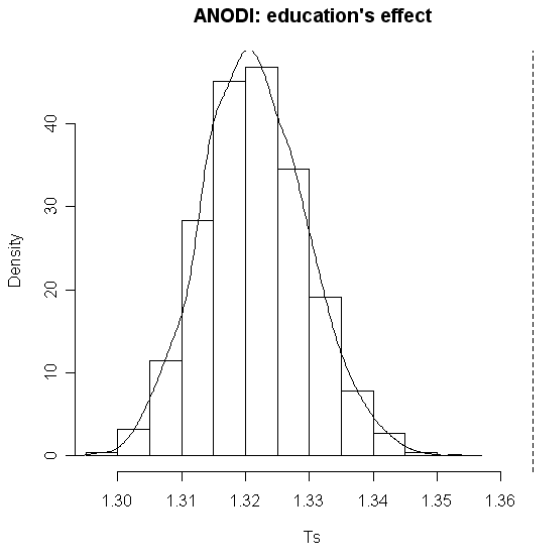

Figure 3.4: Permutation distribution of the statistic $T = B_D/W_D$ based on the categories of "Educ"-covariate of Dutch FFS data. The dashed vertical line marks the observed value of the statistic, T_0.

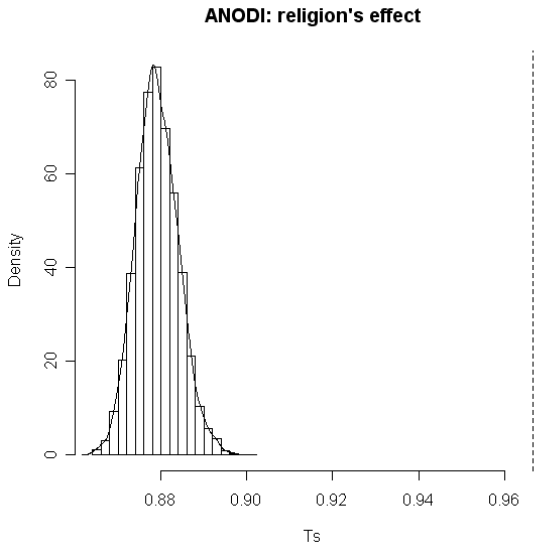

Figure 3.5: Permutation distribution of the statistic $T = B_D/W_D$ based on categories of "Religion"-covariate of Dutch FFS data. The dashed vertical line marks the observed value of the statistic, T_0.

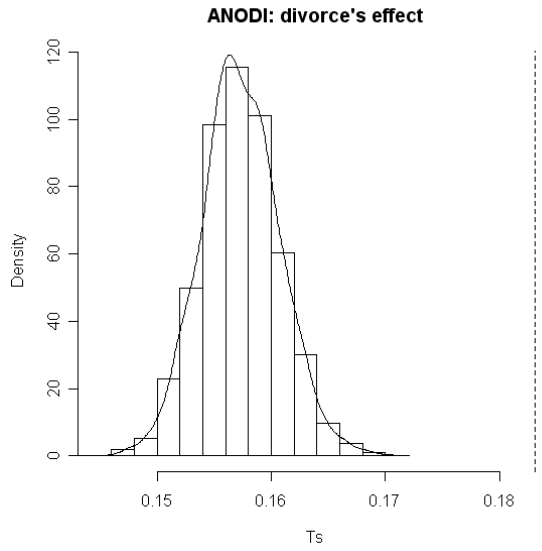

Figure 3.6: Permutation distribution of the statistic $T = B_D/W_D$ based on categories of "Divorce"-covariate of Dutch FFS data. The dashed vertical line marks the observed value of the statistic, T_0.

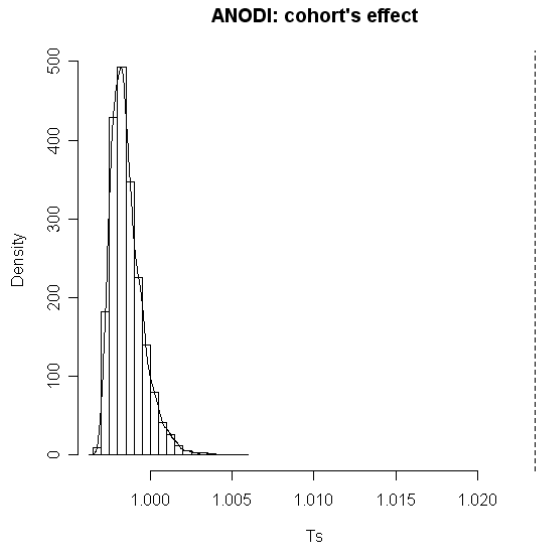

Figure 3.7: Permutation distribution of the statistic $T = B_D/W_D$ based on categories of "Cohort"-covariate of Dutch FFS data. The dashed vertical line marks the observed value of the statistic, T_0.

displaying the sequences after sorting them by the first multidimensional scaling coordinate. On the horizontal axis we place the rank of individuals, while time in months is on the vertical axis. Each state of the sequences is indicated by a specific color: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC. We use this representation for each group of each variable of the explanatory structure. The comparison of the groups' representations allows the graphical inspection of the differences in sequences in such groups. In order to deal with the fact that the sizes of the groups are often very different, we scale the width of the sequences so as to have the same overall span in the horizontal axis. This allows for a better appreciation of the difference and the similarities in the groups. For the parental divorced group we double the width of the sequences because it would be too small to have a clear representation.

Figures 3.8-3.11 show the results, which confirm the results of ANODI test. Several differences appear as long as groups induced by variables are concerned.

- Educ: an increase of the level of education leads to a reduction of the frequency and the duration of the state MC, to an increase of the frequency of the state U and to an increase of individuals who experiences the state S for all the observation period (S/144). These effects are particularly evident in the third level of education, while the first two levels appear more homogeneous.

- Religion: among religious women we notice an increase of the frequency and the duration of the states MC and M, and a reduction of the state U.

- Divorce: parental divorce leads to have a strong reduction of the frequency and the duration of the state M, a strong increase of the proportion of more complex sequences and a mild increase of the state U.

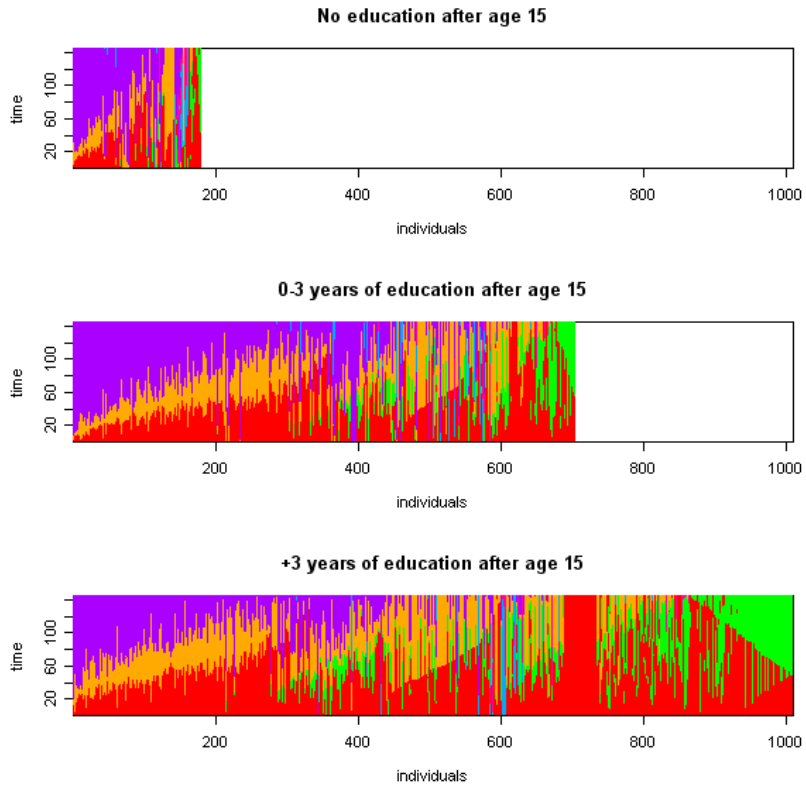

Figure 3.8: Representation of individual sequences of states, separated in education' s level groups, using MDS components ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC.)

- Cohort: the younger cohort is less inclined to experience the state MC but more inclined to the state U. Moreover sequences S/144 are more common in this group.

In the previous part of this work, we noted that the first two groups induced by the variable "Educ" are quite homogeneous. Before proceeding to model selection, we want to understand whether the division into three groups is reasonable or whether it is preferable to merge the first two groups to reduce the number of groups induced by the covariates in the models. In presence of very small groups, such as those induced by divorce, reducing the number of levels of the regressors can help. In order to do this, we apply the ANODI test considering only the part of the distance matrix that

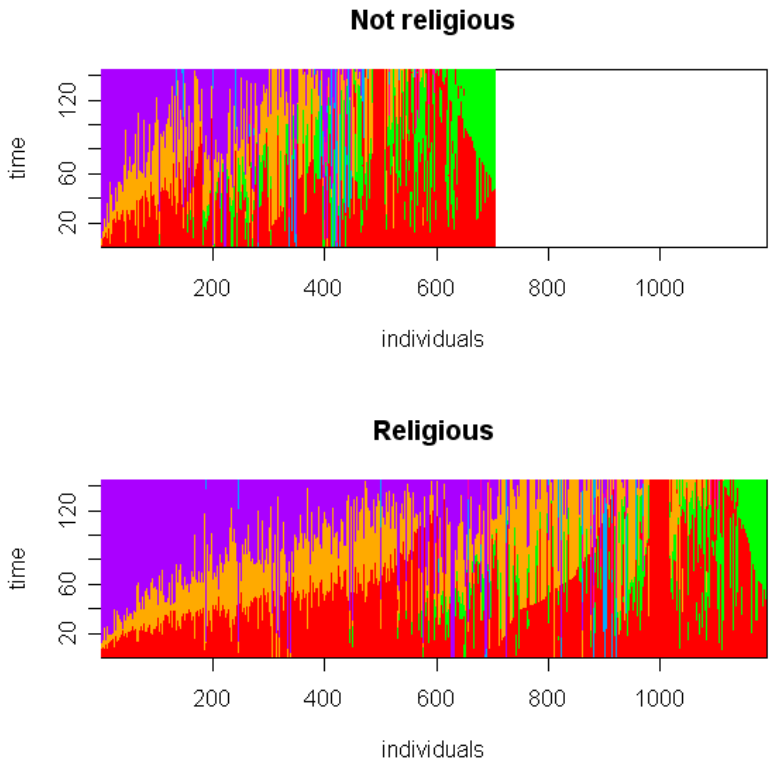

Figure 3.9: Representation of individual sequences of states, separated in religious and not religious women, using MDS components ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC).

concerns these two groups. In this way we test the hypothesis that the two groups are significant to explain the distance between sequences. The result is a p-value equal to 0.33 from which we conclude that the two groups can indeed be merged. As a consequence we use a binary variable for the level of education that distinguishes between women who complete their education before the age of 18 and those who continue to study after that age. The other variables included in the complete model are the same seen above: Religion, Divorce and Cohort.

All the variables are included in the first model. At each step of the selection procedure, each variable is tested for elimination. The p-values of all variables are thus

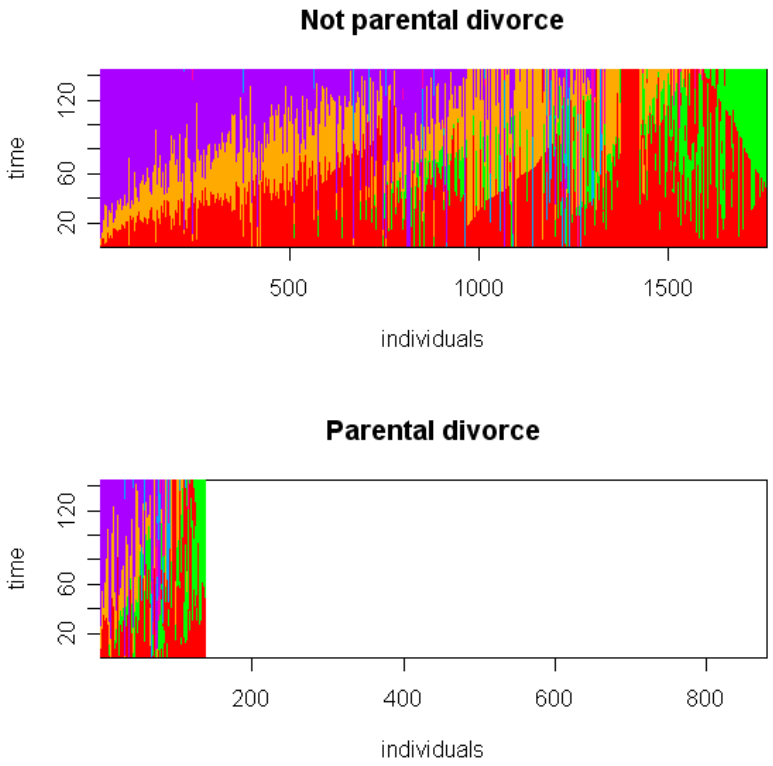

Figure 3.10: Representation of individual sequences of states, separated in groups with or without parents divorced, using MDS components ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC).

calculated, and the variable with the highest (not significant) p-value is removed. The process stops when the p-values corresponding to all the remaining variables are less than 0.05 (other thresholds are possible). Table 3.3 details the process used to arrive at the final model through the evaluation of the partial effects. Based on our analysis results we conclude that the effects of the covariates "Cohort" and "Divorce" are not significant for the determination of the differences between sequences, once the others factors "Educ" and "Religion" are controlled for. The method does not provide a qualitative evaluation of the effect of the covariates, but some indications can be obtained

Complete model			
COVARIATE	F_0^*	**(min(F^*),max(F^*))**	**p-value**
Educ	1.014	(0.969 1.008)	0
Religion	0.909	(0.845, 0.909)	0
Divorce	0.155	(0.138, 0.177)	0.64
Cohort	0.996	(0.974,1.009)	0.82
mod.1: cohort removed			
COVARIATE	F_0^*	**(min(F^*),max(F^*))**	**p-value**
Educ	1.044	(0.973 1.005)	0
Religion	0.923	(0.852 0.910)	0
Divorce	0.157	(0.140 0.176)	0.55
mod.2: divorce removed			
COVARIATE	F_0^*	**(min(F^*),max(F^*))**	**p-value**
Educ	1.045	(0.978 1.001)	0
Religion	0.967	(0.855 0.902)	0

Table 3.3: Selection procedure based on partial effects. Variable are:
Education: education after the age 18 (omitted category: education up to the age 18).
Religion: religiousness (omitted category: not religious).
Divorce: parental divorce (omitted category: no parents divorced).
Cohort: birth cohort 1958-1962 (omitted category: cohort 1953-1957).

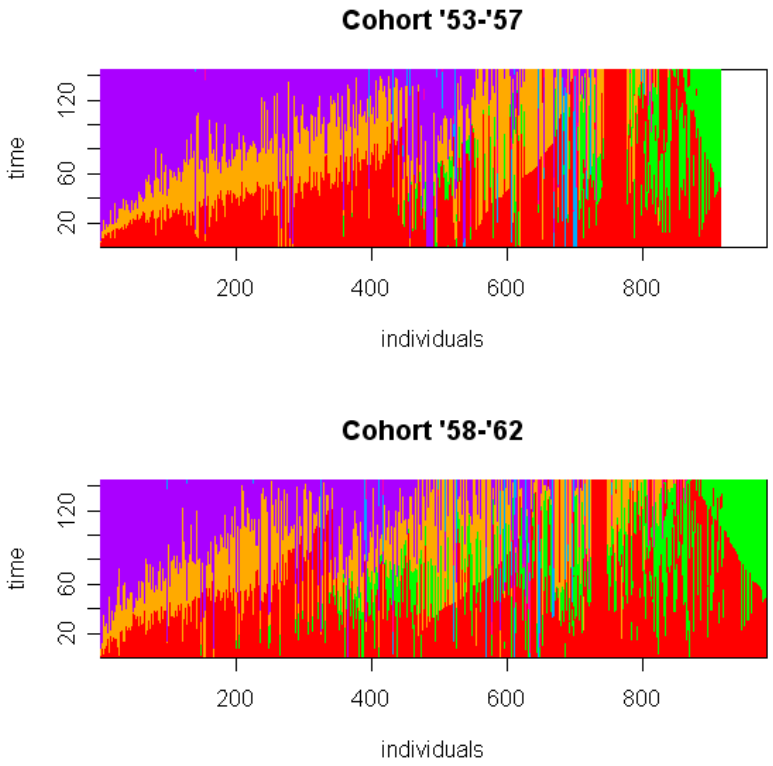

Figure 3.11: Representation of individual sequences of states, separated for cohort birth, using MDS components ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC).

from the graphic analysis described above. We check the effect of the elimination of the tested variable from the model by comparing the representation of the sequence groups induced by the variables in the more complete model with that in the model without the tested variable. In this way we can graphically notice if the further partition leads to a significant change in the structure of the sequences. We exclude from this analysis the variable "Divorce" because the representations obtained with the "parental divorced" group were difficult to interpret due to their reduced sizes. Figures 3.12-3.14 show the graphic representation of the model selection procedure. From Figure 3.11

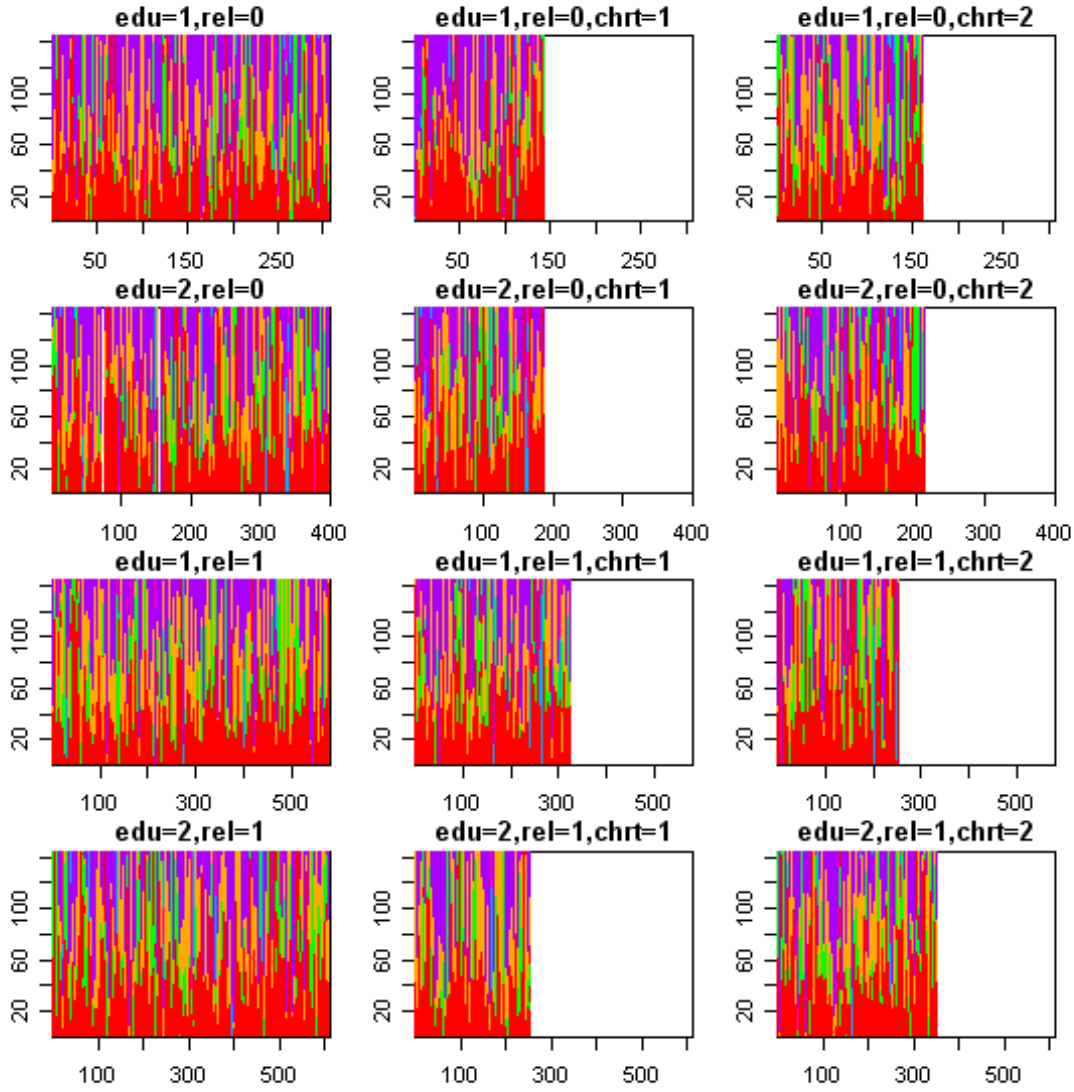

Figure 3.12: Representation of individual sequences of states, separated in groups induced by "Educ" and "Religion" (column 1) and by "Educ", "Religion" and "Cohort" (columns 2-3), using MDS components ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC).

one can notice that once the level of education and the religiousness effects are controlled for, the further partition in cohort groups does not lead to significant differences in the sequence structure. On the contrary, Figures 3.12 and 3.13 demonstrate that

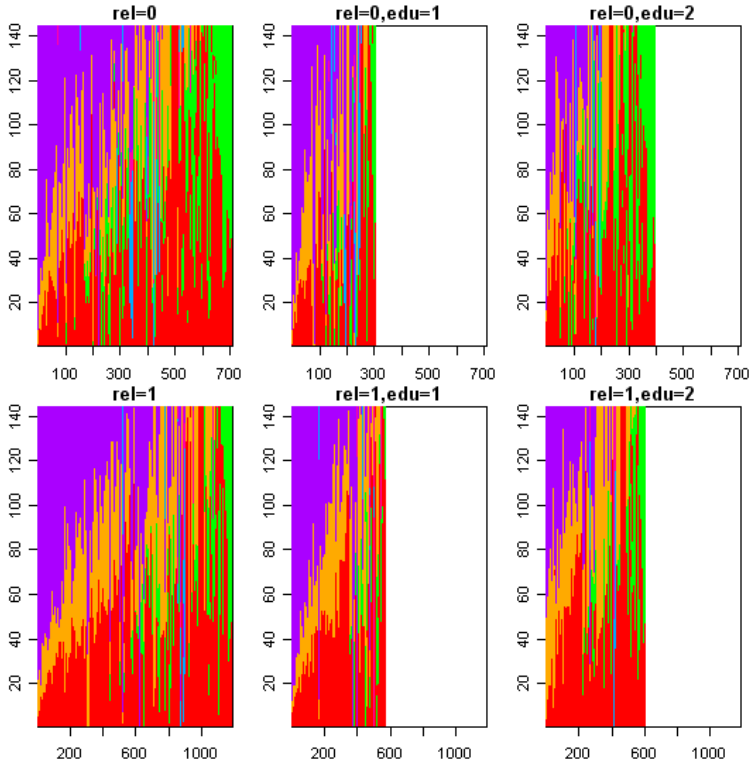

Figure 3.13: Representation of individual sequences of states, separated in groups induced by "Religion" (column 1) and by "Religion" and "Educ" (columns 2-3), using MDS components ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC).

some differences appear as long as groups induced by variables "Educ" and "Religion" are concerned. The more educated women favor the cohabitation while religious ones favor marriage with children. Therefore results of model selection are confirmed by this graphical analysis.

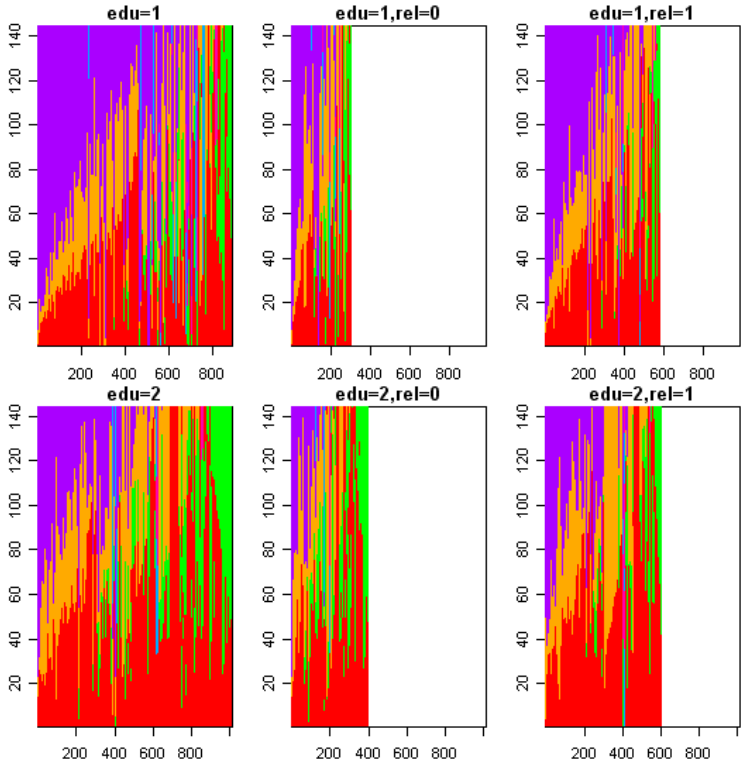

Figure 3.14: Representation of individual sequences of states, separated in groups induced by "Educ" (column 1) and by "Educ" and "Religion" (columns 2-3), using MDS components ordering. Time in months is on the vertical axis and the rank of sequences on the horizontal one (colors indicate different statuses: red = S, orange = M, green = U, sky-blue = SC, violet = MC, pink = UC).

3.6 Conclusions

In this chapter we have introduced an analysis of dispersion (ANODI) approach that allows the evaluation of the effect of one factor on sequence distances, using permutation tests. We applied this technique to sequences of categorical states derived from the FFS data in order to investigate life course trajectories. The method was also used to build a parsimonious model by explaining the relationship between the sequences and the explanatory variables.

The proposed approach adds value to the existing literature from an explanatory point

of view. The methods proposed so far, although they nicely categorize the data, do not give straightforward solutions to the problem of identifying whether there exist a connection between the explanatory structure and life courses as whole (according to the sequence representation). ANODI and the model selection techniques we presented are new procedures in sequence analysis that allow identify which factors affect the differences in the life trajectories and to identify the most parsimonious model to explain the relationship between the distance matrix and the explanatory structure. The results of the substantive analyses illustrate the usefulness of the methods. What makes them particularly attractive is the fact that they exploit all the information contained in the distance matrix, in contrast to existing methods that investigate the explanatory structure starting from a simplified data structure, such as clusters. We provided also a limited simulation study that confirms the efficacy of the ANODI test under a variation of conditions.

The principal limitation of this procedure concerns the interpretation of the estimated covariate effects, even with the information obtainable from the graphical representations. This analysis could however be fundamental in indicating which covariates are significant for the phenomenon being studied.

The proposed model can be extended to different types of model selection methods, for example a forward selection procedure or the Akaike Information Criterion, which takes into account the number of variables in the model.

Chapter 4

A parametric approach to sequence analysis

4.1 Introduction

The description of life courses and, in particular, transition to adulthood, has attracted increasing interest in the literature. A number of methods have been proposed to study trajectories of events with two main goals: to find common patterns among a set of sequences and to study how they are generated. (For a review see Abbott, 1995 or Abbott and Hrycak, 1990). In this chapter we deal with the second objective by proposing a model that generates sequences whose properties resemble those of the original data.

To this aim we model the whole process that generates the paths with a combination of (discrete) time-to-event distribution and transition probabilities. These parametric models are allowed to depend on covariates through generalized logit models. We estimate the transition probabilities and the duration distributions between subsequent transitions from state to state, as well as we compute the probability that a given individual experiences a certain transition status for each combination of characteristics. Our model is part of the broader family of transitional models for the analysis of time sequences. For the study of specific events concerning the transition to adulthood,

event history analysis (Courgeau and Leliévreis, 1992) is one of the principal toolkits of demography. A basic event history problem (Lawless, 2003) can be formulates in terms of a sequence of lifetime variables $T_1, T_2, ...$ that represent the lengths of sojourns (permanences) in a specified sequence of states for an individual. In the sequences T_j, $j = 1, 2, ...$, with $T_j \geq 0$, for a given $j = 1, 2, ...$, the variableT_j represents the same phenomenon for all individuals. Models for $T_{1i}, T_{2i}, ...$ for an individual i, given a vector of covariates Z_i (some of which may be time varying) can be formulated as a sequence of conditional distributions

$$F_j(t|Z_u, t_i^{j-1}) = Pr(T_{ji} \leq t|Z_i, t_i^{j-1}) \tag{4.1.1}$$

where $t_i^{j-1} = (t_{1i}, ..., t_{j-1i})$. In modeling life course transitions, the quantity of fundamental interest is the so-called hazard rate,

$$\lambda_j(t|Z_{ji}, t_i^{j-1}) = lim_{\Delta t \to 0} \frac{Pr(T_{ji} < t + \Delta t|T_{ji} \geq t, Z_{ji}(t), t_i^{j-1})}{\Delta t} \tag{4.1.2}$$

One of the key points is in deciding how to consider the sequence history, t_i^{j-1} and different models were proposed on the basis of this choice: Time-Homogeneity Models, Markov Models and Semi-Markov Models (Meira-Machado, de Ua-lvarez, Cardarso-Surez and Andersen, 2007).

Inference in these models is conducted through a general likelihood. Taking in account for covariates effects and the history of sequence is rather complicated.

It could be useful to write the life course model in terms of one-step transitions. If we indicate with S_t the visited state at time t, $t = 1, 2, ..., T$ and with $f(S_1, S_2, ..., S_T; Z)$ the joint probability mass function of the observed sequence, transitional models use the factorization

$$f(S_1, ..., S_T; Z) = f(S_1; Z)f(S_2|S_1; Z)f(S_3|S_2, S_1; Z)...f(S_T|S_{T-1}, ..., S_1; Z). \tag{4.1.3}$$

Markov Models (see Agresti, 2002, Ch. 11 or Andersen, Borgan, Gill and Keiding, 1993, Ch. 10) and semi-Markov Models (see Example X.1.8 of Andersen, Borgan, Gill and Keiding, 1993 or Andersen, Esbjerg and Sorensen, 2000), which also include explanatory variables Z, are used to compute the probability for movement out of one state into another, the mean permanence time in a given state, the number of individuals in different states at a certain moment (for a specific application of a Markov chain model for the analysis of longitudinal categorical data subject to non-ignorable missingness see Cole, Bonetti, Zaslavsky and Gelber, 2005). Covariates can also explain differences in life course among the individuals.

Note that the link between the model we adopt in this chapter and Markov chain model is strong. Indeed one can be rewritten in form of the other one and vice versa. Let us think to a simple sequence $AAAB$. The probability of observing such sequence can be expressed as a combination of discrete time to event distribution and transition probabilities conditionally on there being a transition:

$$p(AAAB) = p(\text{transition at } 3|A) \cdot p(A \to B|transition),$$

but in an equivalent way as

$$p(AAAB) = p(A) \cdot p(A|A) \cdot p(A|AA) \cdot p(B|AAA) = p(A) \cdot p(A|A)^2 \cdot p(B|A).$$

The advantages deriving from our model over the transitional chain approach are a greater interpretability of the results, due to the direct vision of the effects of the covariates and of the history of the sequence on both the sojourn times and the transition probabilities.

The chapter is structured as follows. In Section 2 we describe the model and the maximum likelihood estimation of the parameters. Section 2 includes also a small simulation

studies of the properties of the model estimators. The Fertility and Family Survey data are appropriate for the study of the transition to adulthood and we fit the model to the Dutch subset of that data in Section 3. In Section 4 we summarize the main findings.

.

4.2 The model

Consider an individual who has visited a total of $r+1$ states, not all necessarily different (i.e. who has experienced r state transitions). Let $\{S_k, k = 1, ..., r+1\}$ indicate the states that the individual experiences, in the order in which they have been visited, and $\{T_k, k = 1, ..., r+1\}$ be the corresponding times spent in the $r+1$ states.
Let Z_k denote a matrix of covariates observed at time k. Indeed these covariates can in principle be time-varying and in the motivating application this is the case. Let β and γ be a vector and a matrix of parameters. We assume the following:

a) T_k follows a geometric distribution with parameter p_k that depends on covariates through the logit link:

$$p_k = \frac{exp(\beta' Z_k)}{1 + exp(\beta' Z_k)} \tag{4.2.1}$$

b) The probability of transitioning from state i to state j at the k-th transition, conditionally on the covariate vector Z_k, is the quantity $P_{ijk} = Pr(S_k = j | S_{k-1} = i, Z_k)$, where $i, j = \{1, ..., J\}$ and $k = 1, ..., K$, with K representing the maximum number of possible transitions, J the total number of states, and $P_{iik} = 0$, $\forall i$, $\forall k$. These transition probabilities are modeled through generalized logit (see for example Agresti, 1990). Using the J-th category as a reference, the logits can be parametrized as

$$P_{ijk} = \frac{exp(\gamma'_{ij} Z_k)}{1 + \sum_{h=(1,...,J-1)-(i)} exp(\gamma'_{ih} Z_k)}. \tag{4.2.2}$$

Each individual's contribution to the loglikelihood depends on the number of (not necessarily different) states visited. For a woman who visits $(r+1)$ states (r transitions) for durations $\{t_k, k = 1, ..., (r+1)\}$, the contribution is:

$$l(\theta|S,T,Z) = \sum_{k=1}^{r}[log(P_{ijk}|Z_k) + log(P(T_k = t_k|Z_k))] + log(P(T_{r+1} > t_{r+1}|Z_{r+1})) \tag{4.2.3}$$

to account for the transitions, the durations in different states and the fact that the last duration is always censored. Here we call $\theta' = (\beta', \gamma')$ the full parameter vector. Let $OT \in \{0, ..., K\}$ be the random variable that counts the number of observed transitions. The parameter θ can be estimated from the observed data on n women by maximizing the observed data log-likelihood

$$l(\theta|S,T,Z) = \sum_{w=1}^{n}(1(OT_w = 0)l_0(\theta|S,T,Z) + ... + 1(OT_w = K)l_K(\theta|S,T,Z)) \tag{4.2.4}$$

where for a woman w who has experienced r transitions we have

$$l_r(\theta|S_w,T_w,Z_w) = \sum_{h=1}^{r}[log(P_{ijh}|Z_{hw}) + log(P(T_h = t_{hw}|Z_{hw}))] + log(P(T_{r+1} > t_{r+1}|Z_{hw})) \tag{4.2.5}$$

and

$$l_0(\theta|S_w,Z_w) = log(P(T_1 \geq 144|Z_{1w})). \tag{4.2.6}$$

to take into account the fact that the censoring is at time 144 months, the end of the data collection period. In order to improve the optimization process (we used the OPTIM function in R with the "BFGS" method, see e.g. Goldfarb, 1970), we provided the R function with the vector of the partial derivatives of the log-likelihood with respect to the parameters:

$$\frac{\partial l_k}{\partial \beta} = \sum_{s=1}^{n} -t_k \frac{exp(\beta' Z_s)}{1 + exp(\beta' Z_s)} + \sum_{r=1}^{k-1}(1 - t_{sr})\frac{exp(\beta' Z_s)}{1 + exp(\beta' Z_s)} Z_s \tag{4.2.7}$$

and, for a given transition from state i to state j

$$\frac{\partial p_{ij}}{\partial \gamma_{ij}} = \sum_{r=1}^{k-1} \frac{-exp(\gamma_{ij}' Z_s)}{1 + \sum_{h=(1,...,J-1)-(i)} \gamma_{ih} Z_s}' Z_s + Z_s \quad (4.2.8)$$

$$\frac{\partial p_{ij}}{\partial \gamma_{il}} = \sum_{r=1}^{k-1} \frac{-exp(\gamma_{il}' Z_s)}{1 + \sum_{h=(1,...,J-1)-(i)} \gamma_{ih} Z_s}' Z_s \quad l \neq j \quad (4.2.9)$$

$$\frac{\partial p_{ij}}{\partial \gamma_{ml}} = 0 \qquad m \neq i \quad l = 1, ..., J \quad (4.2.10)$$

The number of parameters of this model can be very high but some of them can be set to zero to obtain more parsimonious models, as constraints can be applied to the permanence times and to the transition probabilities. Traditional likelihood theory based hypothesis testing can be used to select the "best" model.

We performed a small simulation study of the estimators to examine how well they perform with finite sample size. We simulated sequences of 144 consecutive family life states from a model as described above. For the parameter p of the geometric distribution we included a single covariate, using the parameter values $\beta_0 = -4$ and $\beta_1 = 1$. The covariate was generated from a Bernoulli distribution with parameter 0.4. For the transition probabilities we choose $\gamma_{ij} = (1, -2)$ $i = 1, ..., 6$ $j = 1, ..., 4$. Sample size was 2000. We generated 100 samples and estimated the coverage probability for 80 per cent asymptotic confidence interval corresponding to each parameter. As can be seen in Table 4.1, results indicate generally adequate coverage.

4.3 Analysis of FFS Data

The motivating data for our analysis originate from Dutch Fertility and Family Surveys (FFS).[1]. The dataset contain the retrospective histories of 1897 women between the age of 18 to 30 about childbearing and union formation, collected on a monthly time scale.

[1]For a detailed description of the data and FFS see Section 1.2

Model parameters	Interval coverage
β_0	0.75
β_1	0.81
γ_{111}	0.83
γ_{211}	0.73
γ_{311}	0.85
γ_{411}	0.83
γ_{511}	0.82
γ_{611}	0.83
γ_{121}	0.80
γ_{221}	0.74
γ_{321}	0.78
γ_{421}	0.84
γ_{521}	0.87
γ_{621}	0.78
γ_{131}	0.77
γ_{231}	0.74
γ_{331}	0.78
γ_{431}	0.87
γ_{531}	0.86
γ_{631}	0.86
γ_{141}	0.77
γ_{241}	0.77
γ_{341}	0.78
γ_{441}	0.86
γ_{541}	0.77
γ_{641}	0.80
γ_{112}	0.87
γ_{212}	0.76
γ_{312}	0.79
γ_{412}	0.83
γ_{512}	0.79
γ_{612}	0.82
γ_{122}	0.84
γ_{222}	0.84
γ_{322}	0.79
γ_{422}	0.83
γ_{522}	0.82
γ_{622}	0.79
γ_{132}	0.76
γ_{232}	0.84
γ_{332}	0.84
γ_{432}	0.84
γ_{532}	0.86
γ_{632}	0.80
γ_{142}	0.81
γ_{242}	0.80
γ_{342}	0.75
γ_{442}	0.86
γ_{542}	0.80
γ_{642}	0.79

Table 4.1: 80 per cent confidence interval coverage probabilities based on 100 simulated samples.

Covariates	No.Groups	Frequencies	
		absolute	relative
Educ	3	$n_1 = 181$ $n_2 = 703$ $n_3 = 1009$	$f_1 = 0.10$ $f_2 = 0.37$ $f_3 = 0.53$
Religion	2	$n_0 = 705$ $n_1 = 1188$	$f_0 = 0.37$ $f_1 = 0.63$
Divorce	2	$n_0 = 1754$ $n_1 = 139$	$f_0 = 0.93$ $f_1 = 0.07$
Cohort	2	$n_1 = 912$ $n_2 = 981$	$f_1 = 0.48$ $f_2 = 0.52$

Table 4.2: The Netherlands covariates' description.

We use in the model some informations on individual and family characteristics: level of education, religion belief, parental divorce and birth cohort. All four covariates were used to model both the permanence times and transition probabilities (these covariates are summarized in Table 4.2).

Moreover, we inserted in the model for the permanence times, the age (rescaled and expressed in months) of each individual before she entered that state (Age), as well as the state visited in that moment (variable name Pr-St). Among the covariates that explain the transitions, we also added "Age" and the time spent in the state before the transition (Tval). Note that these covariates change at each visited state.

Of 1897 women, 4 were excluded from this analysis because they are the only ones who experienced transition from S to MC, from S to UC, and from U to MC (the number of observations for the parameter estimates of these transitions was not sufficient for estimation). In addition, transitions from states with children to states without children are not possible and therefore we forced these transition probabilities to zero. The same was done for parameters of other transitions that were never observed. The marginal counts and the relative frequencies of transitions in the observed data are shown in Tables 4.3 and 4.5. Note that because of the structure of the model, the tables refer to transitions conditionally on the fact that a transition did occur. The most frequent transition from M was MC, from U was M, from SC was UC, from UC was SC, from

	S	M	U	SC	MC	UC
S	0	920	911	40	0	0
M	32	0	8	0	1140	0
U	178	554	0	0	0	47
SC	0	0	0	0	19	68
MC	0	0	0	67	0	7
UC	0	0	0	16	52	0

Table 4.3: Number of total transitions in FFS Dutch data.

UC was MC, and from S women opted most frequently for U or M. The sparseness of the data did not allow for the fitting of the model on the various "Children" states, and as a consequence we grouped the last three states (SC, MC, UC) into a unique absorbing state "C". Three sample trajectories are shown in Figure 4.1. The marginal counts and the relative frequencies of the transitions are shown in Tables 4.4 and 4.6. Lastly, Table 4.7 displays the count and the relative frequencies of the visited states in the order in which they have been visited.

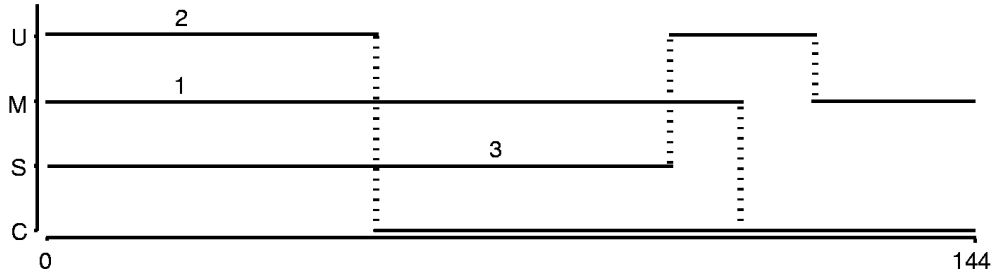

Figure 4.1: Three examples of possible observed life sequences (S: living single without children, M: living married without children, U: living in unmarried cohabitation without children, C: with at least one child. Time (months) is on the horizontal axis.

	S	M	U	C
S	0	920	911	40
M	32	0	8	1140
U	178	554	0	47

Table 4.4: Number of total transitions in FFS Dutch data.

	S	M	U	SC	MC	UC
S	0	0.49	0.49	0.02	0	0
M	0.03	0	0.01	0	0.97	0
U	0.23	0.71	0	0	0	0.06
SC	0	0	0	0	0.22	0.78
MC	0	0	0	0.91	0	0.09
UC	0	0	0	0.24	0.76	0

Table 4.5: Number of transitions (per cent) in FFS Dutch data.

Because the C state is the baseline category, we analysed the effects of the covariates on the odds that individuals have a transition to other states (S or M or U) instead of state C. Initially we fitted a model that included all the covariates and, following a backward selection approach, we sequentially removed the non-significant terms (i.e. $p \geq 0.05$) with likelihood-ratio tests. (For qualitative predictors with more than two

	S	M	U	C
S	0	0.49	0.49	0.02
M	0.03	0	0.01	0.97
U	0.23	0.71	0	0.06

Table 4.6: Number of transitions (percent) in FFS Dutch data.

Visited state	Absolute frequencies				Relative frequencies			
	S	**M**	**U**	**C**	**S**	**M**	**U**	**C**
1st	1781	24	43	45	0.94	0.01	0.02	0.03
2nd	12	937	788	52	0.01	0.52	0.44	0.03
3rd	171	464	14	801	0.12	0.32	0.01	0.55
4th	11	16	107	322	0.02	0.04	0.23	0.71
5th	13	53	5	14	0.15	0.62	0.06	0.16
6th	1	2	5	34	0.02	0.05	0.12	0.81
7th	2	1	–	4	0.29	0.14	–	0.57
8th	–	–	–	2	–	–	–	1

Table 4.7: Visited states' description.

categories we considered the entire variable rather than just one of its dummy thus adding or dropping the entire variable). Table 4.8 details the process used to select the final model, and Table 4.9 contains maximum likelihood estimates of parameters of the final model.

Results from the estimated model suggest that being religious increases strongly the probability of having a transition from S to M with respect to the transition from S to C (Odds Ratio=6.48) and also increases the probability to make transition from M to U with respect to the transition from M to C (Odds Ratio=1.96), but considering the small number of transitions from M to U (8 cases) one should not over-interpret this effect. Moreover, having divorced parents increases the probability of making a transition from S to U with respect to the transition from the same state to C (Odds Ratio=2.44) and decreases the probabilities to make a transition from M to U compared to the one from M to C (Odds Ratio=0.34) and a transition from U to M respect to the transition to C (Odds Ratio=0.34). These considerations are based on keeping the other covariates' values fixed. Therefore parental experience seems to encourage

Model	No.Pars.	Description	chi-square	d.f.	p-value
0	57	Full model	-	-	-
1	55	$\beta_{Educ} = 0$ in Model 0	0.24	2	0.89
2	54	$\beta_{Religion} = 0$ in Model 1	0.7	1	0.40
3	53	$\beta_{Divorce} = 0$ in Model 2	0.72	1	0.40
4	52	$\beta_{Cohort} = 0$ in Model 3	6.14	1	0.01
5	51	$\beta_{Pr-St} = 0$ in Model 3	731.3	2	0
6	52	$\beta_{Age} = 0$ in Model 3	468.26	1	0
7	47	$\gamma_{Tval} = 0$ in Model 3	39.22	6	$6.4 \cdot 10(-7)$
8	47	$\gamma_{Divorce} = 0$ in Model 3	47.5	6	$1.48 \cdot 10(-8)$
9	41	$\gamma_{Educ} = 0$ in Model 3	79.16	12	$5.3 \cdot 10(-15)$
10	47	$\gamma_{Cohort} = 0$ in Model 3	85.44	6	$2.22 \cdot 10(-16)$
11	47	$\gamma_{Age} = 0$ in Model 3	127.08	6	0
12	47	$\gamma_{Religon} = 0$ in Model 3	134.64	6	0

Table 4.8: Summary of sequential likelihood-ratio test to get a final parsimonious model (see text for the definition of the covariates).

less stable relations, such as cohabitation and childless, and to increase the number of divorces.

In addition, the level of education results in statistically significant main effects: most educated women have higher probability of making transition from M to U instead of to C (Odds Ratio=3.32). Also, belonging to the youngest cohort increases this probability (Odds Ratio=2.19). This is not unexpected, since the younger cohort shows higher percentages of women participating in education and prolonged education period. The preference for cohabitation rather than childbearing is probably due to the fact that timing of the first birth and educational level are clearly related. As education postpones the first birth, and indeed the number of childless women aged 30 years has been rising over time. Second, the Dutch society is characterized by the increase of non-marital cohabitation and divorce (Latten, De Graaf, 1997), and the trendsetters are precisely the higher educated women.

Finally, being in state S decreases the probability to make transitions when compared to M and U that are less stable states. Our model estimates that the mean of the

Parameters	**Estimate**	**S.E.**	p**-value**
Permanence			
β_0	-4.329	0.0301	0
β_{Age}	-0.012	0.0006	0
$\beta_{Pr-St=M}$	0.900	0.0410	0
$\beta_{Pr-St=U}$	1.077	0.0458	0
β_{Cohort}	0.079	0.0300	0.018
Transition from S to M			
γ_0	2.799	0.5736	< 0.0001
γ_{Tval}	-0.005	0.0064	0.44
γ_{Age}	-0.054	0.0080	< 0.0001
γ_{Educ2}	0.319	0.5456	0.56
γ_{Educ3}	0.110	0.5274	0.84
$\gamma_{Religion}$	1.869	0.3674	< 0.0001
$\gamma_{Divorce}$	-2.259	0.4213	< 0.0001
γ_{Cohort}	-0.145	0.3413	0.68
Transition from S to U			
γ_0	-4.071	0.8806	< 0.0001
γ_{Tval}	0.013	0.0064	0.04
γ_{Age}	-0.027	0.0086	0.0018
γ_{Educ2}	0.394	0.8540	0.64
γ_{Educ3}	1.546	0.8380	0.06
$\gamma_{Religion}$	-0.253	0.3769	0.5
$\gamma_{Divorce}$	0.890	0.5493	0.10
γ_{Cohort}	0.390	0.3711	0.3
Transition from M to S			
γ_0	2.553	0.7213	0.0004
γ_{Tval}	-0.013	0.0063	0.04
γ_{Age}	-0.015	0.0057	0.01
γ_{Educ2}	-0.107	0.6485	0.86
γ_{Educ3}	0.695	0.6343	0.28
$\gamma_{Religion}$	-0.161	0.3416	0.64
$\gamma_{Divorce}$	-0.515	0.4290	0.22
γ_{Cohort}	-0.384	0.3557	0.28
Transition from M to U			
γ_0	1.652	0.5760	0.004
γ_{Tval}	0.005	0.0063	0.44
γ_{Age}	-0.010	0.0050	0.04
γ_{Educ2}	0.647	0.5446	0.24
γ_{Educ3}	1.201	0.5248	0.02
$\gamma_{Religion}$	0.672	0.3625	0.06
$\gamma_{Divorce}$	-1.083	0.3762	0.004
γ_{Cohort}	0.782	0.3356	0.02
Transition from U to S			
γ_0	-4.657	1.3044	0.0004
γ_{Tval}	0.007	0.0136	0.62
γ_{Age}	-0.0003	0.0130	0.98
γ_{Educ2}	0.567	1.1027	0.60
γ_{Educ3}	-1.088	1.3913	0.44
$\gamma_{Religion}$	-0.400	0.7203	0.58
$\gamma_{Divorce}$	-1.086	2.3363	0.64
γ_{Cohort}	-0.473	0.7413	0.52
Transition from U to M			
γ_0	2.932	0.6849	< 0.0001
γ_{Tval}	-0.018	0.0059	0.002
γ_{Age}	-0.008	0.0052	0.12
γ_{Educ2}	0.398	0.6048	0.50
γ_{Educ3}	0.819	0.5963	0.16
$\gamma_{Religion}$	0.101	0.3168	0.76
$\gamma_{Divorce}$	-1.076	0.4019	0.008
γ_{Cohort}	0.062	0.3342	0.86

Table 4.9: Parameters estimates for the selected model.

permanence time spent is about 76 months in S, 31 in M, and 26 in U. In other words, the estimated expected time spent in M and U is respectively 60% and 65% lower than that spent in S. Note that these permanences might be repeated more than once for some individuals.

For the interpretation of the parameters it could also be useful to consider the multinomial logit model directly in terms of transition probabilities

$$P_{i_k j_k k} = \frac{exp(\gamma'_{i_k j_k} Z)}{1 + \sum_{h=(1,\ldots,J-1)-(i_k)} (exp(\gamma'_{i_k h_k} Z))}. \tag{4.3.1}$$

So, for example, the probability that a transition (when happens) is from "S" to "M" for a 18 years old, religious student, with parents who did not divorce, belonging to the '58-'62 cohort is equal to $8.84 \cdot 10^{-4}$, while for a women not religious but with the same other characteristics that probability is equal to $6.96 \cdot 10^{-3}$. For any typology of women and for any transition it is thus possible to compute the transition probability and compare it with the other ones.

Using the estimated parameter values, the relative probabilities of other more complicated events may be calculated from the model, as it shown at the end of this section in which we consider, for example, the theoretic distribution of the time spent in the first visited state. In alternative to this model-based approach, complicate functionals and events may be estimated by generating from the model a number of sequences, and goodness of fit can be explored by comparing the distribution of observed quantities with those generated by the fitted model. In order to do this, we simulated 1000 samples of 1893 sequences (under the estimated model), censored at 144 months as in the original data.

With the first state visited and the covariates as given, we obtained the times of permanence and the transitions in different states. We compare them with the observed permanence times and observed transitions overall and within strata defined by the

COVARIATES	No	$\bar{T}$ observed	$\bar{T}$ no censoring	$\bar{T}$ censoring
Pr-St=S,Cohort=1	852	51.3	77.0	65.3
Pr-St=S,Cohort=2	929	52.6	71.3	61.9
Pr-St=M,Cohort=1	17	30.8	31.6	31.3
Pr-St=M,Cohort=2	7	32.3	29.3	29.1
Pr-St=U,Cohort=1	15	19.9	26.9	26.8
Pr-St=U,Cohort=2	28	31.8	24.9	24.9

Table 4.10: Comparison between the mean of the observed times in the first state, the mean of times simulated with our model without censoring after 144 months, and with censoring, given each covariates' combination.

COVARIATES	No	observ. $\bar{T}$	simul. $\bar{T}$ no cens	simul. $\bar{T}$ cens	E(T) no cens	E(T) cens
All	1848	51.1	72.3	62.2	-	-
Pr-St$\neq$ M	1824	51.4	72.9	62.6	75.9	64.2
Pr-St=M	24	31.2	30.9	30.9	30.8	30.5
Pr-St$\neq$U	1805	51.7	73.4	63.1	75.9	64.2
Pr-St=U	43	27.7	25.6	25.6	25.8	25.7
Cohort=1	884	50.3	75.3	64.0	75.9	64.2
Cohort=2	964	51.9	69.6	60.6	70.16	60.9

Table 4.11: Comparison between the mean of the observed times in the first state and the mean of times simulated with our model, the expected value estimated, with and without censoring, given each covariate.

covariates. For brevity we report the results only for the permanence times (Tables 4.10-4.13). The other tables are in the Appendix.

For the first visited state, the mean of times of permanence were compared also with the mean of times generated without censoring after 144 months and the expected value of the permanence given each covariate with and without censoring, $E(T|X_i = x_i)$ and $E(T_{cens}|X_i = x_i)$, $i = \{Pr - St = M, Pr - St = U, Cohort\}$. For the second, the third, and fourth we compare only the mean of observed and simulated permanences. Results are shown in Tables 4.10-4.13.

COVARIATES	observed No			simulated No			observed T			simulated T		
	2nd	3rd	4th	2nd	3rd	4th	2nd	3rd	4th	2nd	3rd	4th
Pr-St=S,Cohort=1	4	71	5	5.9	63.6	2.7	21.5	34.6	41.8	66.6	54.4	49.9
Pr-St=S,Cohort=2	8	100	6	8.1	88.5	7.7	26.5	30.7	9.8	65.1	54.5	49.6
Pr-St=M,Cohort=1	542	146	8	467.0	116.9	7.2	44.5	36.9	34.7	38.6	37.2	37.4
Pr-St=M,Cohort=2	395	318	8	356.6	253.6	7.0	48.5	34.7	19.9	36.9	36.4	36.7
Pr-St=U,Cohort=1	273	7	35	248.9	5.0	17.9	37.2	21.4	23.9	33.8	34.6	30.8
Pr-St=U,Cohort=2	515	7	72	455.8	5.7	32.6	36.4	34.0	25.0	32.5	32.8	30.7

Table 4.12: Comparison between the mean of observed and censored simulated permanence times for the second, the third, and fourth visited state, given each covariates' combination.

COVARIATES	observed No			simulated No			observed T			simulated T		
	2nd	3rd	4th	2nd	3rd	4th	2nd	3rd	4th	2nd	3rd	4th
All	1737	649	134	1542.4	533.3	75.1	41.7	34.6	24.9	35.9	41.7	34.6
Pr-St≠M	800	185	118	718.8	162.8	60.9	36.5	31.9	24.6	33.6	53.1	34.0
Pr-St=M	937	464	16	823.6	370.5	14.3	46.2	35.4	27.3	37.9	36.6	37.2
Pr-St≠ U	949	635	27	837.7	522.5	24.6	45.9	34.6	26.1	38.4	41.8	42.4
Pr-St=U	788	14	107	704.7	10.8	50.5	36.7	27.7	24.7	32.9	33.6	30.8
Cohort=1	819	224	48	721.8	185.4	27.7	41.9	35.7	27.5	37.2	43.0	34.4
Cohort=2	918	425	86	820.6	347.9	47.4	41.5	33.7	23.5	34.7	40.9	34.7

Table 4.13: Comparison between the mean of observed and censored simulated permanence times for the second, the third, and fourth visited state, given each covariate.

As these tables and those in the Appendix show, generally the number of transitions from the model are similar to the observed transitions, while permanence times, especially for the first visited state, tend to be overestimated.

This aspect is also evident looking at the estimated survival distributions: to explore the goodness of fit of the model one can compare the survival distributions from simulated sequences (boldface line) with the survival from observed data (confidence intervals are based on the log hazard). Figures 4.2-4.4 display the the survival distributions of the first three permanence times, while Figures 4.5-4.7 show those of the events: first birth, first union (marriage or cohabitation), and first marriage. When examining

these curves one has to keep in mind that the underlying variability of the model-based curves depends on both the fact that we are generating data, and on the sampling variability of the estimated parameters. Therefore the boldface curves, which appear without confidence interval, are actually likely to have quite a variability associated with them.

As mentioned above, figures show that the model tends to overestimate the (first

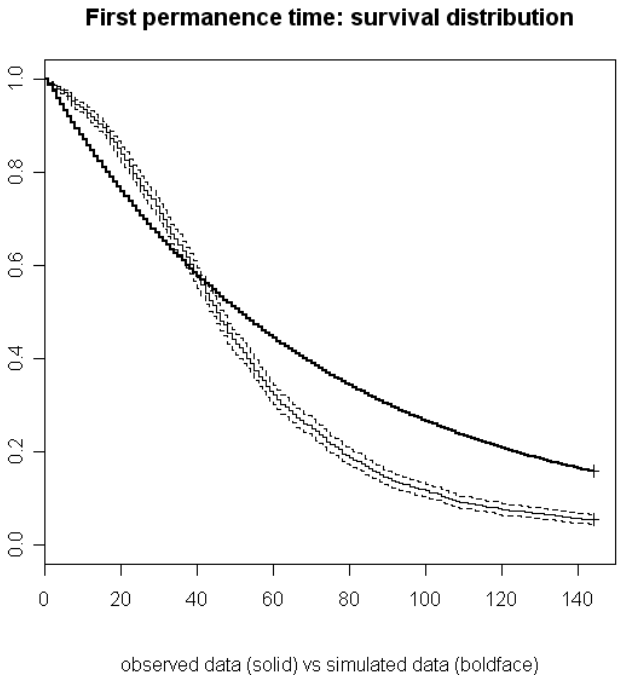

Figure 4.2: Observed (solid) and simulated (boldface) survival distributions for the time in the first visited state.

and third) permanence times. But the model-based curves fit adequately the second permanence time and the first union and first marriage events.

An alternative way (not shown) to interpret the survival distribution results is to compare the data-based curves with the theoretic-model curves of $P(T_k > t_k|Z)$ for any given combination of covariates. For example, given the geometric distribution assumption, it is not difficult to obtain the theoric distribution for the first permanence

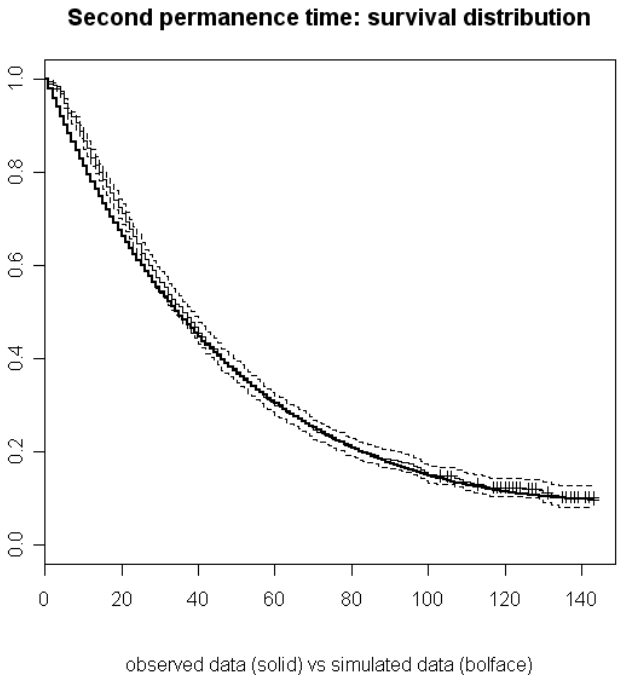

Figure 4.3: Observed (solid) and simulated (boldface) survival distributions for the time in the second visited state

time:

$$P(T_1 > t_1|Z) = \int_Z P(T_1 > t_1|Z) f_Z(Z) dZ \hat{=} \frac{1}{n} \sum_{i=1}^{n} \hat{P}(T_1 > t_1|Z_i) = \frac{1}{n} \sum_{i=1}^{n} (1 - \hat{p}(Z_i))^{t_1} \tag{4.3.2}$$

Improvements in the fit of the model may be possible if one assumes different distributions for the time spent in the states, T_k than the geometric distribution considered here. We replicated the analysis using Weibull distributions, but results were similar and we omit them for brevity. This issue however deserves further analysis using other distributions or combinations of different distributions.

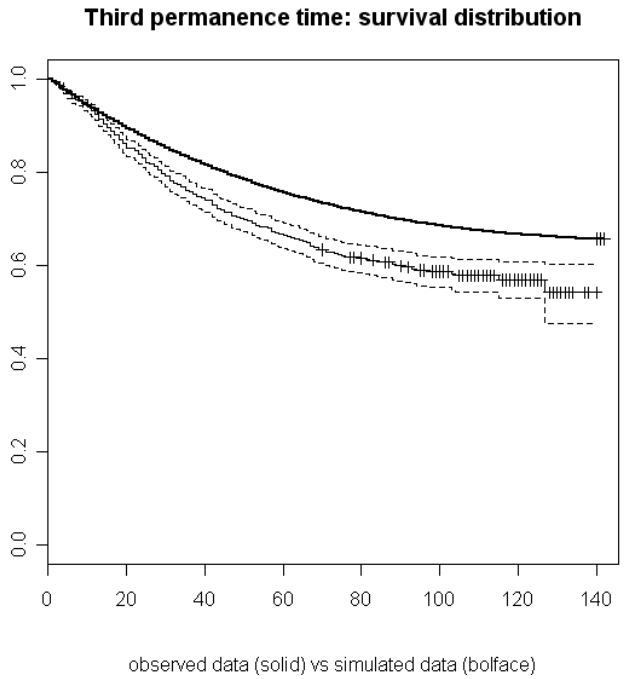

Figure 4.4: Observed (solid) and simulated (boldface) survival distributions for the time in the third visited state

Covariate	***p*-value observed data**	***p*-value simulated data.**
Educ	0	0
Religion	0	0
Divorce	0	0
Cohort	0	0

Table 4.14: Summary table for ANODI-permutation test on Dutch FFS data and on a simulated sample under the estimated model;
Education: education after the age 18 (omitted category: education up to the age 18)
Religion: religiousness (omitted category: not religious)
Divorce: parental divorce (omitted category: no parents divorced)
Cohort: birth cohort 1958-1962 (omitted category: cohort 1953-1957).

4.4 Conclusion and remarks

We have introduced a parametric transitional model to the analysis of time sequences, and applied it to Dutch FFS data. The proposed model accommodates covariate effects (including time-varying covariates), both for the time-to-event distribution and

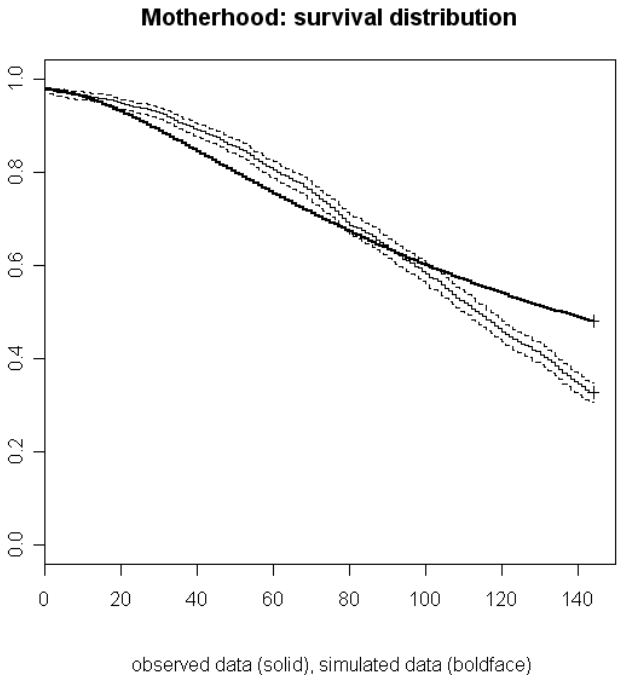

Figure 4.5: Observed (solid) and simulated (boldface) survival distributions for the first birth event

the transition probabilities. A simulation study of a simplified version of the model (including only one not time-varying covariate) was performed to ensure accuracy of the coverage probabilities of the asymptotic confidence intervals of the parameters. From a computational point of view, the model presents several complications, from the delicate identification and fitting to more practical ones, such as data manipulation. We believe that some of the difficulties of fitting these models are related to these aspects. Goodness of fit was assessed empirically by generating data according to the estimated parameters. Summary statistics of the sojourn times and of the frequencies of the transitions among states were compared overall and within strata defined by the baseline covariates. We believe that improvements are possible assuming different distributions for the time spent in the states.

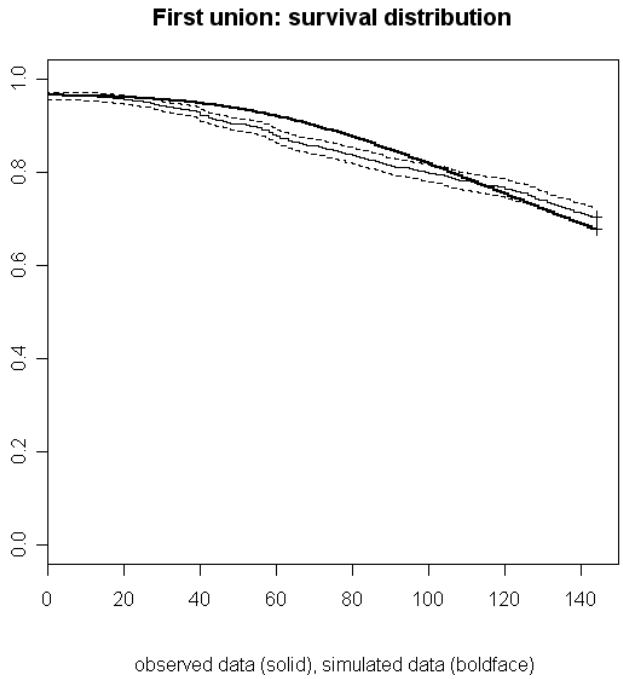

Figure 4.6: Observed (solid) and simulated (boldface) survival distributions for the first union event (cohabitation or marriage)

We compared the results of ANODI procedure presented in Chapter 3 for the observed data with a simulated sample under the estimated model. The simulated sample consists of 1893 women with the same individual and background factors of the observed sample (level of education, religiousness, parental divorce and birth cohort). The method allows the evaluation of the effect of one factor on sequence distance, and it is also used to build a parsimonious model explaining the relationship between the sequences and the explanatory variables. Table 4.14 shows the principal effect of the factors.

Results from the two samples are consistent with each other. For both the original and simulated data all the factors have a significant impact on the distances between cases. Although following different paths, the backward selection model procedure leads to the same model in both samples. "Cohort" and "Divorce" are excluded leading to the

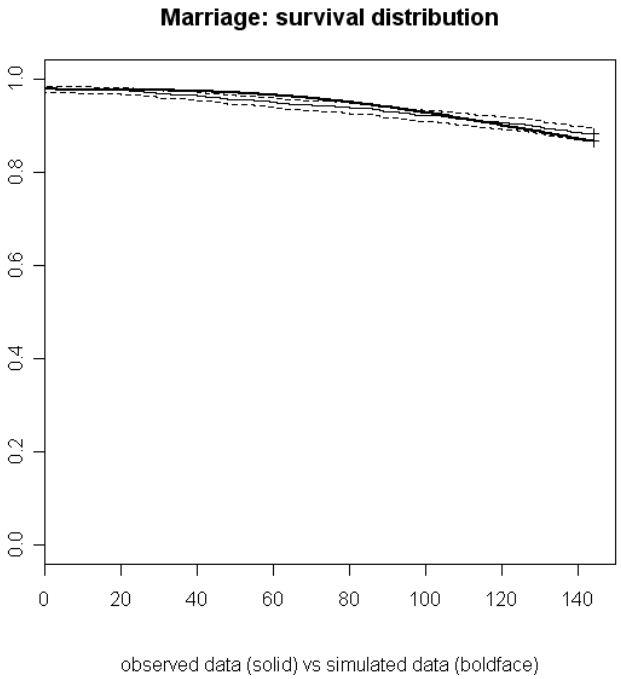

Figure 4.7: Observed (solid) and simulated (boldface) survival distributions for the first marriage event

final model that includes only "Education" and "Religion". Difficulties in fitting duration distribution emerged but ANODI results suggest that the fitted model replicates the structure of sequences quite well.

4.5 Appendix

In order to analyse the transitions of different class of individuals, we compared the observed with simulated absolute frequencies for the most numerous groups induced by the combinations of covariates used in the estimated model (Tables 4.15-4.18).

Educ=2,Religion=1,Divorce=0,Cohort=1		**observed**				**simulated**			
		S	**M**	**U**	**C**	**S**	**M**	**U**	**C**
	S	0	177	49	4	0	155.5	36.8	1.4
	M	4	0	4	175	2.0	0	1.7	129.4
	U	12	34	0	2	5.9	21.2	0	1.7

Table 4.15: Comparison between the observed and simulated transitions for women religious, with education level 2, parents not divorced or separated, and belonging to '53-'57 cohort.

Educ=3,Religion=1,Divorce=0,Cohort=1		**observed**				**simulated**			
		S	**M**	**U**	**C**	**S**	**M**	**U**	**C**
	S	0	150	91	3	0	144.6	72.7	1.8
	M	4	0	0	153	5.7	0	0.3	121.9
	U	19	51	0	5	13.9	34.8	0	2.0

Table 4.16: Comparison between the observed and simulated transitions for women religious, with education level 3, parents not divorced or separated, and belonging to '53-'57 cohort.

Educ=2,Religion=1,Divorce=0,Cohort=2		**observed**				**simulated**			
		S	**M**	**U**	**C**	**S**	**M**	**U**	**C**
	S	0	127	77	0	0	109.2	63.1	1.1
	M	3	0	1	147	2.5	0	1.0	111.5
	U	13	56	0	5	7.0	39.0	0	3.0

Table 4.17: Comparison between the observed and simulated transitions for women religious, with education level 2, parents not divorced or separated, and belonging to '58-'62 cohort.

Educ=3,Religion=1,Divorce=0,Cohort=2		**observed**				**simulated**			
		S	**M**	**U**	**C**	**S**	**M**	**U**	**C**
	S	0	157	179	3	0	133.4	164.0	1.8
	M	4	0	0	190	9.78	0	0.3	155.3
	U	22	119	0	6	24.4	93.4	0	4.8

Table 4.18: Comparison between the observed and simulated transitions for women religious, with education level 3, parents not divorced or separated, and belonging to '58-'62 cohort.

	No observed			No simulated			$\bar{T}$ observed			$\bar{T}$ simulated		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
Age ≤ 50	1848	1037	153	1848.0	901.1	196.8	51.1	44.9	43.8	62.2	36.0	48.4
Age ≥ 51	0	700	496	0	641.7	336.2	-	36.9	31.5	-	35.7	37.7

Table 4.19: Mean of the observed and censored simulated permanence times for two classes of ages at entry into the first, the second and third visited states.

To study the impact of variables "Age" in the permanences sub-model, we compared the observed with the simulated times spent in the first three visited states for the two classes of women, "Age" ≤ 50, "Age" ≥ 51. To study the impact of variables "Age" and "Tval" in the transitions sub-model, we compare the observed with the simulated transitions for the classes, "Age" ≤ 50, "Age" ≥ 51, and for the classes "Tval" ≤ 41, "Tval" ≥ 41. Tables 4.19-4.23 report the results.

Age $\leq$ 50	observed				simulated			
	S	**M**	**U**	**C**	**S**	**M**	**U**	**C**
S	0	913	833	32	0	807.2	741.6	27.7
M	22	0	4	630	32.2	0	4.3	612.1
U	114	305	0	24	108.0	273.2	0	19.4

Table 4.20: Number of observed and simulated transitions for women that have transitions before *51*-th months from start of observation period (22 years and 3 months old).

Age $\geq$ 51	observed				simulated			
	S	**M**	**U**	**C**	**S**	**M**	**U**	**C**
S	0	7	78	8	0	1.3	22.6	2.1
M	10	0	4	510	3.8	0	1.6	261.7
U	64	248	0	23	38.8	146.5	0	14.0

Table 4.21: Number of observed and simulated transitions for women that have transitions after *51*-th months from start of observation period (22 years and 3 months old).

Tval $\leq$ 40	observed				simulated			
	S	**M**	**U**	**C**	**S**	**M**	**U**	**C**
S	0	429	412	24	0	441.8	333.0	15.5
M	15	0	4	741	19.7	0	3.7	612.2
U	134	431	0	34	105.7	320.9	0	20.2

Table 4.22: Number of observed and simulated transitions in which the time spent in the origin state was less than 40 months.

Tval ≥ 41		**observed**				**simulated**		
	S	**M**	**U**	**C**	**S**	**M**	**U**	**C**
S	0	491	499	16	0	366.8	431.2	14.3
M	17	0	4	399	16.4	0	2.2	261.7
U	44	122	0	13	41.1	98.8	0	13.2

Table 4.23: Observed and simulated transitions in which the time spent in the origin state was more than 40 months.

Conclusions

The aim of this thesis was to present new methods for the description and explanation of life course patterns according to their representation as sequences. The thesis consistently applied such methods to the transition to adulthood. The motivating data for our analysis originate from Fertility and Family Surveys (FFS). In order to asses the impact of socio-demographic characteristics on the transition to adulthood, we proposed three new methods which are particularly attractive because they maximize the use of the data, as against previous methods that tried to investigate the explicative structure starting from more simplified data. The previous methods, although they nicely categorize the data, do not give straightforward solutions to the problem of identifying whether or not there exists a connection between the explanatory structure and the life courses. Indeed, we have shown that the multinomial model used by McVicar and Anyadike-Danes (2002), applied to Dutch FFS data, performed quite poorly as far as forecasting is concerned.

The first proposal consists of regression models that use the output distance analysis to detect and predict sub-groups as potential targets for specific public policies. In particular, we analysed adolescent premarital pregnancy adopting an international comparative approach. The goal was to seek individual and origin family factors associated with this risky motherhood behavior. The approach we adopted is to define a template, representing the model long-term single mother, i.e. a women who has had children and has been single during the entire observation period (age 18-30). Then

to use the distance, transformed to the open unit interval (0,1), between this template and the observed sequence for each woman in our sample in a regression model. Results actually suggest that several social and demographic factors influence teenage premarital childbearing.

Moreover we emphasized, by showing that different metrics lead to very different results, the importance of selecting the most appropriate metric with regard to the issue in analysis.

The second proposal was an analysis of dispersion (ANODI) approach that, using permutation test ideas, allows the evaluation of the effect of one factor on sequence distance. To evaluate the relationship between distances and the socio-demographic variables, we adapt the ANOVA approach to the sequence context or, more precisely, to the situation where the unique information about the response variable of interest is synthesized in a dissimilarity matrix. Permutation tests are used to evaluate the significance of an effect. The criterion may also be used to build a parsimonious model explaining the relationship between the sequences and the explanatory variables. The advantage of this method is that it is possible to infer such relationship not from simplified data, i.e. clusters, but directly from the distance matrix which contains more information on the original sequences structure. We also provided a limited simulation study that confirmed the efficacy of the ANODI test under a variety of conditions. Further analyses have to be devoted to have definitive results.

The principal limitation of this procedure concerns the interpretation of the estimated covariate effects: the method indicates which covariates are significant but does not provide a qualitative evaluation of the effect of the covariates. However, some indications can be obtained from the graphical representation of the sequences.

Finally, the third proposal was a parametric model to build the whole process that generates sequences. The method allows us to estimate the effect of the explanatory

variables and also to predict life courses for any given combination of the explicative variables. The explanatory structure was taken into account in the model through generalized logit models. We also performed a simulation study of a simplified version of the model (including only one non time-varying covariate) to ensure the accuracy of the coverage probabilities of the asymptotic confidence intervals of the parameters. For this kind of models, handling goodness of fit is difficult and it depends on the correctness of the parametric assumptions. We investigated goodness of fit empirically by generating data according to the estimated parameters. Summary statistics of the permanence times and of the frequencies of the transitions among states were compared overall and within strata defined by the baseline covariates. Difficulties in fitting duration distribution emerged but we believe that improvements are possible assuming different distributions for the time spent in the states. Moreover, to evaluate the goodness of fit of the model we compared the results of the ANODI procedure for the observed data with a simulated sample based on the estimated model. Results from the two samples are consistent with each other. For both the original and simulated data all the factors have a significant impact on the distances between cases. Although following different paths, the backward selection model procedure leads to the same model in both samples. Therefore ANODI results suggest that the fitted model replicates the structure of sequences quite well. Parametric models do require careful work to be used due to the delicate identification and fitting, and, despite their potentialities, this might make their use not as widespread.

This thesis provides several techniques for the analysis of the sequences of states, from graphical to parametric to nonparametric. Although the techniques are presented as distinct procedures, these techniques can be used sequentially or one to support the others, in order to identify which factors are important for the phenomenon being studied and the qualitative effects of such factors. We linked sequence analysis to other

conventional statistical methods (MDS, cluster analysis, ANOVA, regression models. We are at an early stage and there is scope for additional work in this area.

Bibliography

Abbott, A. (1995). Sequence analysis: New Methods for Old Ideas. *Annual Review of Sociology*, **21**, 93-113.

Abbott A., Tsay A. (2000). Sequence Analysis and Optimal Matching Methods in Sociology. Review and Prospect. *Sociological Methods & Research*, **29(1)**, 3-33.

Abbott, A. (2000). Reply to Levine and Wu. *Sociological Methods & Research*, **29**, 65-76.

Abbott, A., Forrester, J. (1986). Optimal Matching Methods for Historical Sequences. *Journal of Interdisciplinary History*, **15**, 471-494.

Agresti, A. (2002). *Categorical Data Analysis*. Wiley.

Andersen, P. K., Borgan, O., Gill, R. D., Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer Series in Statistics.

Andersen P. K., Esbjerg S., Sorensen T. I. A. (2000). Multistate models for bleeding episodes and mortality in liver cirrhosis. *Statistics in Medicine*, **19**, 587-599.

Aassve, A., Billari, F. C., Piccarreta, R. (2004). Sequence Analysis of BHPS Life Course Data. In: *New Development in Classification and Data Analysis*,

eds. M. Vichi, P. Monari, S. Mignani, and A. Montanari, (pp.275-284). Heidelberg, Springer-Verlag.

Bardone, A. M., Moffitt, T. E., Caspi A., Dickson, N., Silva, P. (1996). Adult mental health and social outcomes of adolescent girls with depression and conduct disorder. *Development and Psychopathology*, **8**, 811-829.

Billari, F. C. (2001). Sequence analysis in demographic research and applications. *Canadian Studies in Population*, **28**, 439-458.

Billari, F. C., Fürnkranz, J., Prskawetz, A. (2006). Timing, sequencing, and quantum of life course events: a machine learning approach. *European Journal of Population*, **22**, 37-65.

Billari, F. C., Piccarreta, R. (2001). Life courses as Sequences: an experiment in clustering via monothetic divisive algorithms. In S. Borra, R. Rocci, M. Vichi, M. Schader (Eds.). *Advances in Classification and Data Analysis*. Springer, Berlin, 351-358.

Billari, F. C., Piccarreta, R. (2005). Analyzing Demographic Life Courses through Sequence Analysis. *Mathematical Population Studies*, **12**, 81-106.

Blossfeld, H. P., Rohwer, G. (2001). *Techniques of Event History Modeling. New Approaches to Casual Analysis*. Lawrence Erlbaum Associates.

Bonetti, M., Pagano, M. (2004). The interpoint distance distribution as a descriptor of point patterns, with application to spatial disease clustering. *Statistics in medicine*,**24(5)**, 753-773.

Bonetti, M., Salford, G. (2007). Parametric transitional models for the

analysis of time sequences: two examples. *Proceedings S. Co. 2007 Modelli complessi e metodi computazionali intensivi per la stima e previsione.* Venezia, 6-8 September 2007.

Capaldi, D. M., Crosby, L., Stoolmiller, M. (1996). Predicting the time of the first sexual intercourse for at-risk adolescent males. *Child Development*, **67**, 344-359.

Cole, B. F., Bonetti, M., Zaslavsky, A. M., Gelber R. D. (2005). A Multistate Markov Chain Model for Longitudinal, Categorical Quality-of-Life Data Subject to Non-ignorable Missingness, *Statistics in Medicine*, **24**, 2317-2334.

Cooksey, E. C. (1990). Factors in the Resolution of Adolescent Premarital Pregnancies. *Demography*, **27**, 207-218.

Courgeau, D., Leliévre, É. (1992). *Event History Analysis in Demography.* Clarendon Press, Oxford.

DeLamater, J. (1981). The Social Control of Sexuality. *Annual Review of Sociology*, **7**, 263-290.

Edington E. S., Onghena, P. (2007). *Randomization tests.* Chapman and Hall/CRC.

Elzinga, C. H. (2003). Sequence Similarity: A Non-Aligning Technique. *Sociological Methods & Research*, **32(1)**, 3-29.

Elzinga, C. H. (2005). Combinatorial Representations of Token Sequences. *Journal of Classification*, **21(1)**, 87-118.

Elzinga, C. H. (2006). Sequence Analysis: Metric Representation of Categorical Time Series. *Sociological Methods & Research*, under revision.

Elzinga, C. H., A.C. Liefbroer (2007). De-standardization of Family-Life Trajectories of Young Adults: A Cross-National Comparison Using Sequence Analysis. *European Journal of Population*, **32**, 225-250 .

Esping-Andersen, G. (1990). *Three Worlds of Welfare Capitalism*. Polity Press, Cambridge.

Esping-Andersen, G. (1999). *Social Foundations of Postindustrial Economies*. Oxford University Press.

Ferrari, S., Cribari-Neto, F. (2004). Beta Regression for Modeling Rates and Proportions. *Journal of Applied Statistics*, **7**, 799-815.

Festy, P., Prioux, F. (2002). *An Evaluation of the Fertility and Family Surveys Project*. United Nations, New York.

Goldfarb (1970). A Family of Variable-Metric Methods Derived by Variational Means. *Mathematics of Computation*, **24(109)**, 23-26.

Halpin, B., Chan, T. W. (1998). Class Careers as Sequences: An Optimal Matching Analysis of Work-Life Histories. *European Sociological Review*, **14(2)**, 111-130.

Hanson, S. L., Myers D. E., Ginsburg A. L. (1987). The Role of Responsibility and Knowledge in Reducing Teenage Out-of-Wedlock Childbearing. *Journal of Marriage and Family*, **49**, 241-256.

Hardin, J. W., Hilbe, J. M. (2007). *Generalized Linear Models and Extensions, 2nd Edition.* Stata Press, Texas.

Hoffert, S. (1987). Social and economic consequences of teenage childbearing. In *Riskingthe Future: Adolescent Sexuality, Pregnancy and Childbearing (Vol. 2)* (pp. 123-144). S. Hoffert and C. Hayes. Washington, DC: National Academic Press.

Howard, B. K., Powell, M. A. (2002). *Effects of Family Structure, Education and Religion on Contraceptive Decisions by Women in their Twenties.* Thesis (M.A.), University of Nebraska at Omaha.

Jessor, R., Jessor S. (1977). *Problem Behaviors and Psychosocial Development: A Longitudinal Study of Youth.* New York, Academic Press.

Jobson, J. D. (1992). *Applied Multivariate Data Analysis.* Springer.

Kiernan K (1999). Childbearing outside marriage in Western Europe. *Population Trends*, **98**, 11-20.

Kowalski, J., Pagano, M. and DeGruttola, V. (2002). A Non-Parametric Test of Gene Region Heterogeneity Associated with Phenotype. *Journal of the American Statistical Association*, **97**, 398-408.

Latten, J., De Graaf, A. (1997). *Fertility and Family Surveys in Countries of the ECE Region. Standard country report. The Netherlands.* UNITED NATIONS, New York and Geneva.

Lattin, J. M., Green, P. E., Douglas Carroll, J. (2002). *Analyzing Multivariate Data.* Duxbury Pr.

Lawless, J. F. (2002). *Statistical Models and Methods for Lifetime Data, 2nd edition.* Wiley.

Levine J. H. (2000). But What Have You Done for Us Lately? Commentary on Abbott and Tsay. *Sociological Methods & Research*, **29(1)**, 34-40.

Lior, O., Piccarreta, R. (2007). *Predicting transitions from school to work careers. A comparison of proposals.* BSc dissertation.

McGullagh, P., Nelder, J. A. (1989). *Generalized Linear Models, 2nd edition.* Chapman and Hall, New York.

Meira-Machado, L., de Ua-lvarez, J., Cardarso-Surez, C., Andersen, P. K. (2007). Multi-state models for the analysis of time to event data. *Research Report 07/1, Department of Biostatistics, University of Copenhagen.*

McVicar, D., Anyadike-Danes, M. (2002). Predicting successful and unsuccessful transitions from school to work by using sequence methods. *Journal of the Royal Statistical Society A*, **165**, 317-334.

Michael, R. T., Tuma, N. B. (1985). Entry into Marriage and Parenthood by Young Men and Women: the Influence of Family Background. *Demography*, **22**, 515-544.

Moore, K. A., Morrison, D. R., Green, A. D. (1997). Effects on the children born to adolescent mothers. In R. Maynard (eds) *Kids having kids: economic costs and social consequences of teen pregnancy* (pp. 145-180). Washington, DC: Urban Institute Press.

Mouw, T. (2006). Sequences of Early Adult Transitions: A Look at Variability and

Consequences. In R.A. Settersten, F.F. Furstenberg Jr. and R.G. Rumbaut (eds) *On the Frontier of Adulthood: Theory, Research and Public Policy* (pp. 256-291). Chicago: University of Chicago Press.

O'Connor, T. G., Thorpe, K., Dunn, J., Golding, J. (1999). Parental divorce and adjustment in adulthood: Findings from a community sample. *Journal of Child Psychology and Psychiatry*, **40**, 777-789.

Papke, L. E., Wooldridge, J. M. (1996). Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates. *Journal of Applied Econometrics*, **11**, 619-632.

Pesarin, F. (2001). *Multivariate permutation tests: with applications in biostatistics*. Wiley.

Piccarreta, R., Billari, F. C. (2005). *Predicting Work and Family Trajectories.* Annual Meeting of the Population Association of America, Philadelphia.

Piccarreta, R., Billari, F. C. (2007). Clustering Work and Family Trajectories using a Divisive Algorithm. *Journal of the Royal Statistical Society.* Forthcoming.

Plotnick, R. D., Butler, S. S. (1991). Attitude and Adolescent Nonmarital Childbearing: evidence form the National Longitudinal Survey of Youth. *Journal of Adolescent Research*, **6**, 470-492.

Plotnick, R. D. (1992). The Effect of Attitudes on Teenage Premarital Pregnancy and its Resolution. *American Sociological Review*, **57(6)**, 800-811.

Rohwer, G., Ptter, U. (2005). *TDA User's manual.* Ruhr-Universität Bochum, Bochum.

Sankoff, D., Kruskal, J.B. (1983). *Time warps, string edits and macromolecules*. Reading, Ma: Addison Wesley.

Schlich, R., Axhausen, K. (2003). Habitual travel behaviour: Evidence from a six-week travel diary. *Transportation*, **30(1)**, 13-16.

Smithson, M., Verkuilen, J. (2006). A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables. *Psychological Methods*, **11(1)**, 54-71.

Stovel, K. and Bolan, M. (2004). Residential Trajectories: The Use of Sequence Analysis in the Study of Residential Mobility. *Sociological Methods & Research*, **32**, 559-598.

Szwarc, S., Bonetti, M. (2006). Modelling menstrual status during and after adjuvant treatment for breast cancer. *Statistics in Medicine*, **25**, 3534-3547.

Tanfer, K., Cubbins L. A., Brewster, K. L. (1992). Determinants on Contraceptive Choice Among Single Women in the United States. *Family Planning Perspective*, **24(4)**, 155-161+173.

Venables, W. N., Ripley, B. D. (2002). *Modern Applied Statistics with S. Fourth Edition*. Springer.

Widmer, E., Levy, R., Pollien, A., Hammer, R., Gauthier, J.-A. (2003). Entre standardisation, individualisation et sexuation : une analyse des trajectoires personnelles en Suisse. *Swiss Journal of Sociology*, **29**, 35-67.

Woodward L., Fergusson D. M., Horwood, L. J. (2001). Risk Factors and Life Processes Associated with Teenage Pregnancy: Result of a Prospective Study

from Birth to 20 Years. *Journal of Marriage and the Family*, **63(4)**, 1170-1184.

Wu, L. L. (2000). Some Comments on "Sequence Analysis and Optimal Matching in Sociology: Prospect and Review". *Sociological Methods & Research*, **29 (1)**, 41-64.

Wu, L. L., Martinson, B. C. (1993). Family structure and the risk of premarital birth. *American Sociological Review*, **58**, 210-232.