UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche

SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE

CICLO XXII

ESSAYS ON SEQUENCE ANALYSIS FOR LIFE COURSE TRAJECTORIES

Direttore della Scuola: Ch.ma Prof.ssa ALESSANDRA SALVAN

Supervisore: Ch.mo Prof. GIANPIERO DALLA ZUANNA

Co-supervisore: Ch.mo Prof. FRANCESCO C. BILLARI

Co-supervisore: Ch.mo Prof. FRANK F. FURSTENBERG

Dottorando: NICOLA BARBAN

31 Luglio 2010

Contents

List of Figures

List of Tables

Abstract

The thesis is articulated in three chapters in which I explore methodological aspects of sequence analysis for life course studies and I present some empirical analyses. In the first chapter, I study the reliability of two holistic methods used in life-course methodology. Using simulated data, I compare the goodness of classification of Latent Class Analysis and Sequence Analysis techniques. I first compare the consistency of the classification obtained via the two techniques using an actual dataset on the life course trajectories of young adults. Then, I adopt a simulation approach to measure the ability of these two methods to correctly classify groups of life course trajectories when specific forms of "random" variability are introduced within pre-specified classes in an artificial datasets. In order to do so, I introduce simulation operators that have a life course and/or observational meaning. In the second chapter, I propose a method to study the heterogeneity in life course trajectories. Using a non parametric approach, I evaluate the association between Optimal Matching distances and a set of categorical variables. Using data from the National Longitudinal Study of Adolescent Health (Add-Health), I study the heterogeneity of early family trajectories in young women. In particular, I investigate if the OM distances can be partially explained by family characteristics and geographical context experienced during adolescence. The statistical methodology is a generalization of the analysis of variance (ANOVA) to any metric measure. In the last chapter, I present an application of sequence analysis. Using family transitions from Wave I to Wave IV of Add-health, I investigate the association between life trajectories and health outcomes at Wave IV. In particular, I am interested in exploring how differences in timing, quantum and order of family formation transitions are connected to self-reported health, depression and risky behaviors in young women. Using lagged-value regression models, I take into account selection and the effect of confounding variables.

Riassunto

La tesi è articolata in tre sezioni distinte in cui vengono afffrontati sia aspetti metodologici che analisi empiriche riguardanti l'analisi delle sequenze per lo studio del corso di vita. Nel primo capitolo, viene presentato un confronto tra due metodi olistici per lo studio del corso di vita. Usando dati simulati, si confronta la bontà di classificazione ottenuta con modelli di classi latenti e tecniche di analisi delle sequenze. Le simulazioni sono effettuate introducendo errori di tipo stocastico in gruppi omogenei di traiettorie. Nel secondo capitolo, si propone di studiare l'eterogeneità nei percorsi di vita familiare. Usando un approccio nonparametrico, viene valutata l'associazione tra le distanze ottenute tramite l'algoritmo di Optimal Matching ed un insieme di variabili categoriche. Usando i dati provenienti dall'indagine *National Longitudinal Study of Adolescent Health (Add-Health)*, si studia l'eterogeneità nei percorsi di formazione familiare di un campione di giovani donne statunitensi. La metodologia statistica proposta è una generalizzazione dell'analisi della varianza (ANOVA) . Nell'ultimo capitolo, si presenta un'applicazione dell'analisi delle sequenze per dati longitudinali. Usando i dati sulla transizione alla famiglia dalla prima alla quarta rilevazione nell'indagine Add-Health, vengono studiate le associazioni tra transizioni familiari e diversi indicatori di salute. In particolare, viene studiato come alcune caratteristiche legate alle transizioni familiari (*timing, quantum, sequencing*) siano associate allo stato generale di salute, depressione e comportamenti a rischio. La selezione e l'effetto di variabili confondenti sono prese in considerazione nell'analisi.

Aknowledges

I would like to express my sincere gratitude to my supervisor, Professor Gianpiero Dalla Zuanna from the Department of Statistical Sciences, University of Padua. His understanding, encouraging and personal guidance have been of great value for me throughout these years. I am deeply grateful to my co-supervisor, Professor Francesco Billari from Università Bocconi for his constant support throughout this work. I owe my most sincere gratitude to Professor Frank F. Furstenberg. This thesis would not have been possible unless his support and suggestions. I warmly thank Professor Michael J. White, for his valuable advice and friendly help during my staying at Brown University. I would also like to thank the Academic board of the PhD school and the faculty member of the Department of Statistical Sciences at University of Padua. I owe my deepest gratitude to all the Dondena researchers and the administrative staff. The welcome and friendship that I encountered at the center has been of great value to me. I am indebted to my many of my PhD colleagues for their support and friendship during these years. I am particularly thankful to my friends Vanna Albieri and Nadia Frigo. I would also like to thank the PhD students in sociology and demography at Brown University, University of Pennsylvania and the participants at the IUSSP summer schools in Lund (2008) and Rostock (2009).

I would like to show my gratitude to the administrative staff at the Department of Statistical Sciences in Padua. Also, I would like to thank Patricia Miller, the administrator of the network on transitions to adulthood at University of Pennsylvania, for her support during my visiting period at University of Pennsylvania. Last, I would like to express my deepest gratefulness to Elisabetta and my family for their continuous support and encouragement during all these years.

This research uses data from Add Health, a program project directed by Kathleen

Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis. This thesis would not been possible without the help of all these people. All errors are my own.

1. Introduction

1.1. Overview and Section

The thesis is articulated in three chapters in which I explore methodological aspects of sequence analysis for life course studies and I present some empirical analyses. During the last decade, there has been an increasing interest in the "holistic" approach to life course analysis. Rather than focusing on the timing of an event, (e.g childbirth, marriage/cohabitation, living parental home) the holistic approach focuses on the trajectories, i.e. the whole part of life course. The aim of the two strategies is different. In a event-based strategy the timing of an event is the variable of interest usually as related to covariates that can be either constant or varying over time. The standard set of statistical techniques is event history analysis, also known as (multivariate) survival analysis. The holistic approach considers the joint evolution over time of events in life course. The variable of interest is not the timing of an event (or a set of events), but rather the entire trajectory. Life course trajectories can be analyzed by representing the original data, i.e. each individual's life course, as a sequence of states. Each sequence is then considered as a whole in the input of statistical analysis. Consequently, studying individual sequences allows to focus not only on the timing of a single event but also on the occurrence, the ordering and the synchronization between different life dominions. This approach is particularly effective to study complex period of the life course such as the transition to adulthood or family formation (Billari, 2005).

In the first section of the thesis, I study the reliability of two holistic methods used in life-course methodology. Using simulated data, I compare the goodness of classification of Latent Class Analysis (LCA) and Sequence Analysis (SA). Although the two methods come from very different statistical tradition, they both have been applied to

recognize typical patterns in life course trajectories. Sequence Analysis is a pure algorithmic technique used to measure the "social distance" between individuals (Abbott and Tsay, 2000). There is no parameter estimation and the resulting output is a matrix of dissimilarities. Distances are then used as input in data reduction techniques, mainly clustering. Latent Class Analysis refers to an unobserved discrete mixture distribution (Hagenaars and McCutcheon, 2002). The estimates can be interpreted as the contribution of every observed variable on defining the class. Both the methods allow to aggregate individuals in homogeneous groups. Doing so, the researcher can simplify the complexity of the life course defining typical pattern of transitions. It is not clear, however, how reliable are the two methods. Are the two methods equivalent? Under what conditions do they produce different results? In the case of life course analysis, we are interested in variations in timing, quantum and sequencing of demographic events. I investigate the reliability of LCA and SA under random variations of life trajectories. I propose a simulation approach to test the robustness of the two methods. I simulate an hypothetical dataset with different groups of family transitions and I introduce stochastic disturbances to test the classification power of the two techniques. Furthermore, I test the consistency of the classification obtained via the two techniques using an actual dataset on the life course trajectories of young adults. The results contribute on the one hand to outline the usefulness and robustness of findings based on the classification of life course trajectories through LCA and SA, on the other hand to illuminate on the potential pitfalls of actual applications of these techniques.

In the second chapter, I propose a method to study the heterogeneity in life course trajectories. Using a non parametric approach, I evaluate the association between the dissimilarity matrix obtained by Optimal Matching algorithm (OM) and a set of categorical variables. Using data from the National Longitudinal Study of Adolescent Health (Add-Health), I study the heterogeneity of early family trajectories in young women. In particular, I investigate if the OM distances can be partially explained by family characteristics and geographical context experienced during adolescence. The statistical methodology is a generalization of the analysis of variance (ANOVA) in the case of semi-metric and non-metric measures. This method has been introduced in ecology by

Anderson (2001b) and McArdle and Anderson (2001) to ecosystem analysis and it has been used by Zapala and Schork (2006) in order to evaluate genetic relations. Since OM distances cannot be assumed normally distributed, I use a permutation approach to assess the statistical significance of the association tests. In this section, I present both univariate and multivariate association tests.

In the last chapter, I present an application of sequence analysis to longitudinal data. Using family transitions from Wave I to Wave IV of Add-health, I investigate the association between life trajectories and health outcomes at Wave IV. In particular I am interested in exploring how differences in timing, quantum and order of family formation transitions are connected to self-reported health, depression and risky behaviors in young women. Using lagged-value regression models, I take into account selection and the effect of confounding variables. Previous studies on health and marital status (Harris et al., 2010; Koball et al., 2010; Wood et al., 2007) focused on changes that occur in marital status rather than the entire trajectory. Sequence analysis allows to analyze the development of life course in order to evaluate if characteristics such as the complexity and the order of life sequences have an impact on later outcomes. Last, I present six typologies of life-trajectories obtained using cluster analysis and I explore the association with health outcomes.

1.2. Main Contributions of the Thesis

- I propose a method to simulate life course trajectories. Starting from an artificial dataset of homogeneous life courses, I propose a series of sequence operators in order to introduce heterogeneity in timing, quantum and sequencing in life course sequences.

- Using simulated data, I compare the goodness of classification of latent class models and sequence analysis for life trajectories.

- I compare the performances of different distance measures for sequence analysis: Optimal Matching algorithm (OM) and Longest Common Subsequences (LCS)

- I propose a method for studying the heterogeneity of life course trajectories. I introduce a method for the analysis of variance using Optimal Matching dissimilarities. The statistical significance is assessed using permutation tests.

- I apply sequence analysis techniques to the National Longitudinal Sample of Adolescent health (Add-Health) in order to investigate the association between family formation trajectories and health from a life course perspective.

2. Classifying life course trajectories: a comparison of latent class and sequence analysis

2.1. Introduction

In recent years, there has been a significantly growing interest in the holistic study of life course trajectories, i.e. in considering whole trajectories as a unit of analysis, both in a social science setting and in epidemiological and medical studies. A particular focus of such research has been the classification of individuals according to life course trajectories, so to develop typical classes, or groups, of trajectories. This chapter contribute to this line of research by assessing the robustness and consistency of the findings obtained using two of the most widespread approaches to such problem, latent class analysis (LCA from now onwards) and sequence analysis (SA from now onwards).

The two techniques, LCA and SA, come from different statistical background. Sequence Analysis, in its various specifications, is based on algorithmic, or data mining, approaches aimed at making use of measures of dissimilarity, or distance, between individual trajectories (see, e.g., Abbott, 1995; Abbott and Tsay, 2000; Billari and Piccarreta, 2005; Elzinga, 2006; Brzinsky-Fay and Kohler, 2010). The SA approach is fully nonparametric, and the standard output of the first step of SA analyses is a matrix of dissimilarities. In the second step, SA-based dissimilarity matrices are then used as inputs in data reduction techniques, mainly cluster analysis or multidimensional scaling. Groups obtained via data reduction can be used, in a third step, in subsequent analyses, e.g. on the determinants or consequences of life course trajectories. Latent Class Analysis, in its various specifications, is based on a probabilistic modeling approach, with a finite mixture distribution as

the data generating mechanism (see, e.g., Hagenaars and McCutcheon, 2002; Lin et al., 2002; Reboussin et al., 2002; Beath and Heller, 2009; Bruckers et al., 2010; Pickles and Croudace, 2010). The underlying hypothesis in LCA models is that individuals belong to a finite number of classes (i.e., the values of a categorical variable) that cannot be observed. The estimating procedure aims at estimating the probability of class membership for each trajectory based on observed data via, usually, a likelihood function. LCA can also be embedded in more complex structural models, where the determinants and consequences of trajectories are included in the model, or life course trajectories are seen in parallel with other processes. Estimates are commonly obtained through an EM algorithm. In terms of classification, LCA also provides the contribution of every observed variable on the definition of classes.

In the social sciences, the analysis of life course trajectories has been applied to elicit typical pathways in the transition to adulthood, professional careers, family and fertility, criminal careers. Using either LCA or SA techniques, individuals are assigned to homogeneous classes that are interpreted as representing typical behaviors (Aassve et al., 2007; McVicar and Anyadike-Danes, 2002; Blair-Loy, 1999; Macmillan and Eliason, 2003; Amato et al., 2008; Nagin and Tremblay, 2005; Roeder et al., 1999; Tremblay et al., 2004; Groff et al., 2010). The resulting distribution in groups can be used to test a specific theory or to compare cohorts, subpopulations or the same population across time and/or space (Billari, 2001; Widmer and Ritschard, 2009). Furthermore, class membership can be used as an explanatory variable for further analyses (McVicar and Anyadike-Danes, 2002; Mouw, 2005; Billari and Piccarreta, 2005; Amato et al., 2008). Sequence analysis has also been used in geographical and mobility studies focusing on transitions that occur not only in time, but also in space. The resulting trajectories represent a set of transitions that individuals experience across time in different locations in space. For example, a SA approach has been used to describe the trajectories of tourists choice behavior (Bargeman et al., 2002; Shoval and Isaacson, 2007), or to classify individuals based on their mobility and daily-activity patterns (see e.g. Wilson, 2001; Schlich and Axhausen, 2003; Stovel and Bolan, 2004; Wilson, 2008; Saneinejad and Roorda, 2009; Vanhulsel et al., 2010)

In biostatistics and epidemiology, most applications make use of LCA or related models.

LCA models are used to identify typical patterns in the evolution of health status during life course and to analyze their determinants (see e.g. Hayford, 2009; Dunn et al., 2006; Harrison et al., 2009; Bruckers et al., 2010; Croudace et al., 2003). Other studies focus on the link between health or behavioral trajectories and later outcomes during life course (Hamil-Luker and O'Rand, 2007; Lajunen et al., 2009; Berge et al., 2010; Savage and Birch, 2010; Haviland et al., 2007). Despite SA techniques were first used in genetics and biostatistics to compare DNA sequences, there are no applications of SA in the study of the evolution of health trajectories during the life course. This is partially motivated by the fact that these studies generally focus on the evolution across time of continuous variables, while SA techniques are generally used to describe trajectories of discrete states. Nevertheless, a large array of medical applications can be described as a sequence of discrete states. For example, the evolution of BMI across life course can be described in categories (e.g. underweight, normal weight, overweight or obese status) using suitable thresholds. Also, SA methods may be used to describe the occurrence and persistence of particular health status such as hypertension, depression or physical limitations.

For what follows, this chapter will particularly focus on the event-based interpretation of holistic approaches to the analysis of life courses. Within this interpretation, the aim of holistic methods is to study simultaneously the *timing* of events in the life course (when do events happen?, e.g. when do individuals experience their first sexual intercourse or smoke their first cigarette), their *sequencing* (in which order do events happen?, e.g. do individuals have a child prior to marriage or stop smoking before the birth of a child), and their *quantum* (how many events happen?, e.g. how many births do they have) (Billari, 2005).

In the remainder of this chapter, I compare the performance LCA and SA and test their consistency. In particular, I focus on the use of LCA and SA as devices to obtain classes of individual life course trajectories. After a brief introduction and review of the relevant literature, I compare the consistency of the classification obtained via the two techniques using an actual dataset on the life course trajectories of young adults. Then, a simulation approach is adopted to measure the ability of these two methods

to correctly classify groups of life course trajectories when specific forms of "random" variability are introduced within pre-specified classes in an artificial datasets. In order to do so, I introduce simulation operators that have a life course and/or observational meaning. The results obtained contribute on the one hand to outline the usefulness and robustness of findings based on the classification of life course trajectories through LCA and SA, on the other hand to illuminate on the potential pitfalls of actual applications of these techniques.

2.2. Life course trajectories as categorical time series

Life course trajectories can be described as the observation, over the course of an individual's time (i.e. age), of a number of events (i.e. life events) triggering a change in a corresponding number of categorical states. The approach used in the analyses can however, without loss of generality, be extended to states that are measurable on a quantitative scale (e.g. systolic blood pressure level, income) over discrete time units. It can also be used to represent the life course of units other than individuals (e.g., households, organizations, institutions, ...).

The concept of trajectory derives from the interdisciplinary systematization of the life course paradigm proposed by Elder (1985), in which life course trajectories usually refer to the joint occurrence of events in multiple life domains. For example, one may want to have a representation of the evolution of union status, childbearing and work history. Trajectories can be analyzed by representing the original data, i.e. each individual's life course, as a sequence of states. Each individual i can be associated to a variable s_{it} indicating her/his life course status at time t. As one can assume that s_{it} takes a finite number of values, trajectories can be described as categorical time series. In other terms, trajectories can be represented as strings or sequences of characters, with each character denoting one particular state. The state-space, (i.e the alphabet from which sequences are constructed) has a finite number of elements and represent all the possible states that an individual can take in each time period. For instance, a woman who is single for 12 months since the start of our observation (e.g., age 18), then starts a cohabitation lasting

5 months and then marries and remains married for 7 months can be described as follows:

$$SSSSSSSSSSSCCCCCMMMMMMM$$

In this case, the state-space has 3 values (S=single; M=married; C=cohabiting).

More formally, let us define a discrete-time stochastic process $S_t : t \in T$ with state-space $\Sigma = \{\sigma_1, \dots \sigma_K\}$ with realizations s_{it} with $i = 1 \dots n$. The life course trajectory of the individual i is described by the sequence $s_i = \{s_{i1} \dots s_{iT}\}$.

For practical reasons, a more compact representation of sequences, which we shall use later on, involves counting the repetitions of a state, which in the former example becomes as follows:

(S,12)-(C,5)-(M,7)

Life course sequences $\{s_{i1} \dots s_{iT}\}$ can be alternatively represented by a series of vectors $\{\mathbf{y_{it}}, \dots \mathbf{y_{iT}}\}$ where the K categories of s_{it} are represented by $M = K-1$ binary variables. This representation is particularly useful in the latent class framework, where the series of binary observations are included in the model through a logistic link.

I now briefly review the use of Latent Class Analysis and Sequence Analysis in the study of life course trajectories.

2.3. Latent Class Analysis of life course trajectories

Latent Class Analysis (LCA) is a statistical technique used (also) to classify individuals based on a set of categorical outcomes (Lazarsfeld and Henry, 1968; Goodman, 1974; McCutcheon, 1987; Clogg, 1995; Hagenaars and McCutcheon, 2002). The underlying assumption of LCA is that individuals belong to classes that are unobserved (latent), but for which observed data provide adequate information on class membership through a likelihood function. When data are collected longitudinally, the use of LCA is usually defined "latent trajectory modeling" or "longitudinal latent class analysis" (Vermunt, 2008b; Beath and Heller, 2009; Collins and Wugalter, 1992).

In the LCA framework, it is convenient to represent the life course trajectory as a series of binary vectors indicating the simultaneous occurrence of states in different life domains. Let us assume that there are i subject, $j = 1, \ldots, M$ life domains, $c = 1, \ldots, C$ classes and $t = 1, \ldots, T$ periods. The conditional likelihood for each subject is:

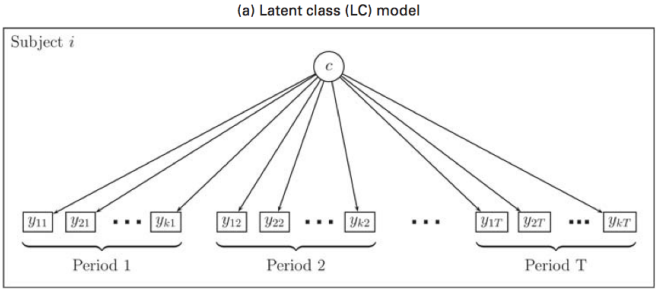
$$P(y_{i11\cdot}, \ldots, y_{iMT}|c_i = c) = \prod_{t=1}^{T} \prod_{j=1}^{M} \pi_{cjt}^{yijt}(1 - \pi_{cjt})^{1-y_{ijt}},$$

where π_{cjt} is the probability of jth outcome =1 at time t for class c, constrained to be between zero and one by transformation through, for example, the logistic scale.
Summing over the classes, weighted by η_c, one obtains the marginal likelihood:

$$P(y_{i11}, \ldots, y_{iMT}) = \sum_{c=1}^{C} \eta_c P(y_{i11}, \ldots, y_{iMT}|c_i = c)$$

LCA assumes that the structure of correlation between observed variables is completely explained by latent factors. This condition is called "conditional indipendence", that is $P(y_{i11\cdot}, \ldots, y_{iMT}|c_i = c) \perp\!\!\!\perp P(y_{i11\cdot}, \ldots, y_{iMT}|c_i = d)$ with $d \neq c$ (Espeland and Handelman, 1989; Hagenaars, 1988; Uebersax, 1999). The longitudinal structure of the model can be represented by figure 2.1.

Figure 2.1.: Latent class structure for longitudinal data, (Beath and Heller, 2009)



The principal drawback of using standard LCA for longitudinal data is that these models do not take in consideration the time correlation between variables. The same variable measured in different time periods is, in fact, considered independent. In the recent years, various forms of correction have been proposed to adjust for temporal correlation between

observations, mainly including a random effect in the model (Vermunt, 2008a; Beath and Heller, 2009; Hadgu and Qu, 1998; Vermunt, 2003). In later analyses, I refer to the more standard version of LCA applied to longitudinal data.

2.4. Sequence analysis and Optimal Matching

Sequence analysis is a family of algorithm based techniques used to quantify distances between categorical time series. Optimal Matching algorithm (OM) is the most known technique that has been applied to social science. The development of OM started in the seventies and the technique has been described in details by Kruskal (1983). Basically, OM expresses distances between sequences in terms of the minimal amount of effort, measured in terms of edit operations, that is required to change two sequences such that they become identical. A set that is composed of three basic operations to transform sequences is used: $\Omega = \{i, \delta, \sigma\}$, where i denotes *insertion* (one state is inserted into the sequence), δ denotes *deletion* (one state is deleted from the sequence) and σ denotes *substitution* (one state is replaced by another state). To each of these elementary operations $\omega_k \in \Omega$, a specific cost can be assigned, $c(\omega_k)$. If K basic operations must be performed to transform one sequence into another the transformation cost can be computed as $c(\omega_1, \ldots \omega_K) = \sum_{k=1}^{K} c(\omega_k)$.

A specific cost can be assigned to each operation, and the total cost of applying a series of edit operations can be computed as the sum of the costs of single operations. The distance between two sequences can thus be defined as the minimum cost of transforming one sequence into the other one. Hence, the resulting output is a symmetric matrix of pairwise distances that can be used for further statistical analysis, mainly multivariate analysis. Optimal Matching is a family of dissimilarity measures derived from the measure originally proposed in the field of information theory and computer science by Vladimir Levenshtein (Levenshtein, 1965). Abbott (1995) adapted OM to social science assigning to three elementary operations different costs, based on the social differences between states (Lesnard, 2006). The choice of the operations' costs determines the matching procedure and influences the results obtained. This is a major concern about the use

of this technique in social sciences (Wu, 2000). A common solution for assessing the substitution costs is to use the inverse of the transition probability, in order to assign higher costs to the less common transitions (Piccarreta and Billari, 2007).

2.4.1. Sequence-based alternatives to Optimal Matching Algorithm

The use of OMA in the analysis of life course trajectories has often been criticized. (for a recent review see Brzinsky-Fay and Kohler, 2010; Aisenbrey and Fasang, 2010).

First, it is difficult to attribute a sociological meaning to the sequence operations (Lesnard, 2006). In biology the three edit operations used in OM are of little theoretical relevance since there is no resemblance with bio-chemical processes. However, differently from biological sequences, social sequences are time referenced. Therefore, the edit operations in social sequences imply modifications in the time scale. In particular, insertion and deletion operations warp time in order to match identically coded states but occurring at different moments in their respective sequences. On the other hand, substituting two events conserve the original time scale of events without warping time. A simple solution to avoid *indel* operations is to use the Hamming distance (Hamming, 1950). The Hamming distance measures the minimum number of substitutions required to change one string into the other.

Second, the choice of costs is a major concern on the use of OM for social sciences because their arbitrariness and the weak link to theory. Critics argue that the resulting distances are meaningless from a sociological point of view (Levine, 2000). In the case in which there is no a clear ranking between the different states, the definition of cost is necessarily arbitrary. A common practice is to set constant costs independent to the states that are substituted. This is equal to set $c(i) = c(\delta)$ and $c(\sigma) = 2c(\delta)$. Using this approach, $c(i)$ is a scaling factor, and the dissimilarity between two sequences is proportional to the (minimum) number of operations that are needed to transform one into another, with double weight given to substitution. The reason for setting $c(\sigma) = 2c(\delta)$ is that, in a constant cost framework, substitution is equivalent to a deletion followed by an insertion. Alternatively, it is possible to adopt a data-driven approach, i.e. using substitution costs that are inversely proportional to transition frequencies (Piccarreta

and Billari, 2007). Consider two states, a and b. Let $N_t(a)$ and $N_t(b)$ be the number of individuals experiencing respectively a and b at time t, and $N_{t,t+1}(a,b)$ be the number of individuals experiencing a at time t and b at time $t+1$. The transition frequency from a to b is

$$p_{t,t+1}(a,b) = \frac{\sum_{t=1}^{T-1} N_{t,t+1}(a,b)}{\sum_{t=1}^{T-1} N_t(a)} \tag{2.1}$$

The cost of substituting a for b is $c(\sigma;a,b) = c(\sigma;b,a) = 2 - p_{t,t+1}(a,b) - p_{t,t+1}(b,a)$ if $a \neq b$. This cost specification takes into account the occurrence of the events weighting more those transitions that are less frequent. A possible critic is that transitions at different age are qualitatively different. For this reason, Lesnard (2006) proposes a modification of the Hamming distance using dynamic costs. The "Dynamic Hamming Distance" (DHD) is based on time-varying substitution costs $c_t(\sigma;a,b)$.

Third, it is not clear how to treat missing data and censoring among sequences. In fact, unequal sequence length due to censoring should not contribute to distance between sequences. A common practice is to restrict the analysis to sequences of the same length in order to avoid distortions due to comparing sequences of different length. Elzinga (2006) proposes different measures for categorical time series that are valid for sequences of different length and do not require cost specification. The basic idea is to compare the number of common subsequences of two sequences in order to asses a similarity measures. A subsequence is a sequence that can be derived from another sequence by deleting some elements without changing the order of the remaining elements. For example, ABD is a subsequence of $ABCDE$. Remarkable subsequences are the prefix and the suffix of a sequence, that are, respectively, the first (last) k elements of a sequence. Elzinga (2006) reviews in details different distance measures based on subsequences. The basic idea is that two sequences are very similar if they have in common long subsequences. In this way, the length of common subsequences can be used as an indicator of the similarity of two strings. Suitable measures based subsequences are: the longest common subsequence (LCS); the longest common prefix (LCP) and the longest common suffix (RLCP). The theoretical basis of these measures come from information science and their great advantage is that the researcher does not need to specify any operation costs.

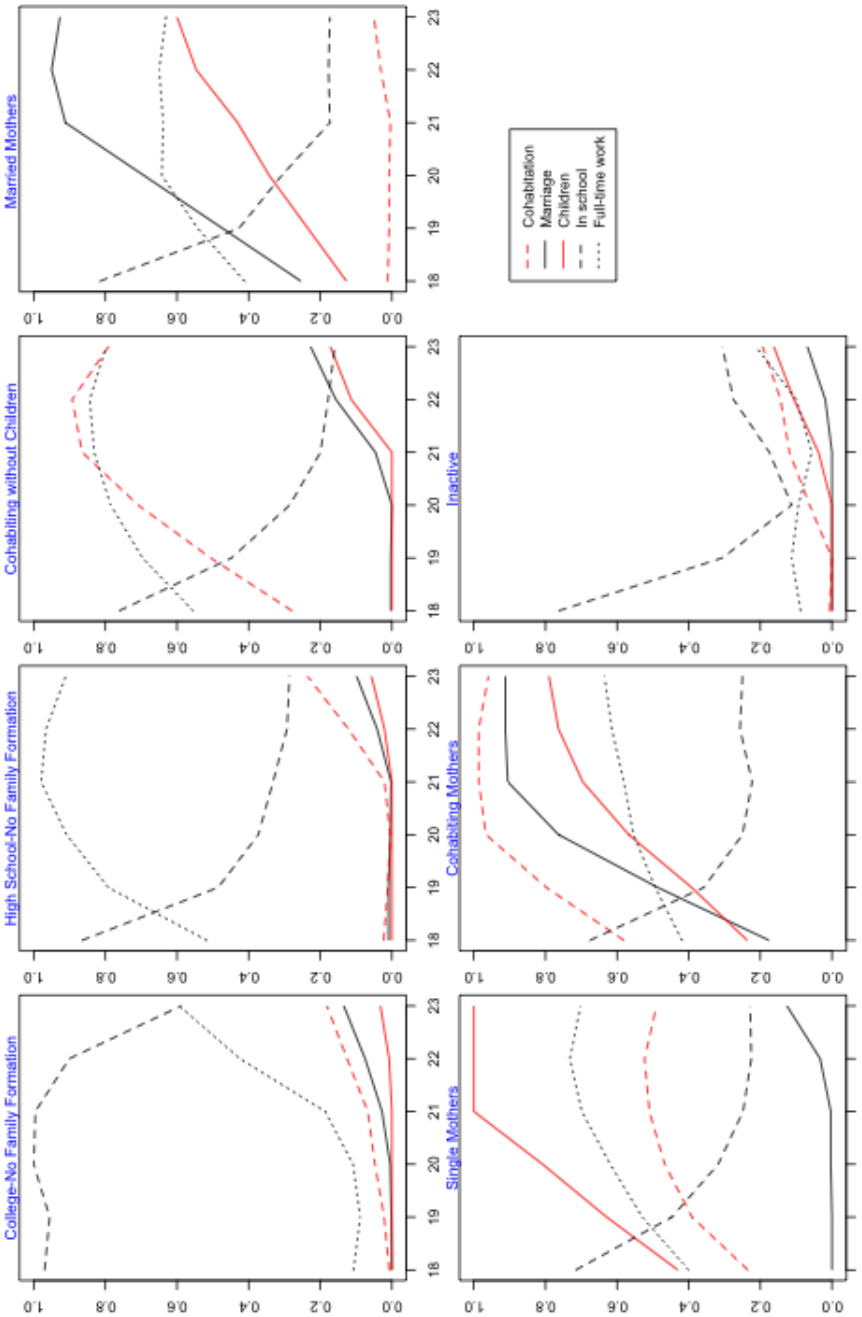
Other solutions that have been proposed rely on OM with some modifications. For example, Hollister (2009) and Gauthier et al. (2009) analyze different cost specification, while Halpin (2010) proposes a modified version of the algorithm where OMs elementary operations are weighted inversely with episode length.

2.5. The consistency of LCA and SA: an example using real life course data

One of the main challenges of studying the life course is the complexity of life course data (Giele and Elder, 1998). It is a common practice in life course analysis to identify sensible periods of the life course using a set of different markers coming from different life domains. For instance, transition to adulthood can be described with five life course transitions: finishing school, beginning full-time employment, entering a non-marital cohabitation, becoming a parent, and getting married. The fact that these transitions can occur in different orders and at different ages yields to an enormous number of possible combinations. To study the diverse experiences of transition to adulthood, it is necessary to reduce the number of pathways to a manageable number. Amato et al. (2008) propose to use latent class analysis to create family formation pathways for women between the age of 18 and 23. Input variables include cohabitation, marriage, parenthood, full-time employment, and school attainment. Data ($n = 2,290$) come from Waves I and III of the National Longitudinal Study of Adolescent Health (Add Health). The analysis revealed seven latent pathways: college- no family formation (29%), high school– no family formation (19%), cohabitation without children (15%), married mothers (14%), single mothers (10%), cohabiting mothers (8%), and inactive (6%). Figure 2.2 shows the estimates of a latent class model.

Would a sequence analysis lead to the same results? The first possible test is to run a sequence analysis with the same data and compare the groups obtained by the two methods. Family formation trajectories can be described by the joint occurrence of the five variables described above. The resulting sequence is 6 period long and the state-space is composed $2^5 = 32$ elements resulting from the combination of the possible states. It

Figure 2.2.: Latent class representation of early family formation. Women 18-23 years old. Add-health, (Amato et al., 2008)



follows that the number of possible sequences is 32^6. To compare the LCA solution with sequence analysis, I calculated the dissimilarity matrix using different distances: OM with transition costs; Longest Common Subsequence (LCS); OM with constant costs;

Dynamic Hamming Distance (DHD); Longest Common Prefix (LCP); Longest Common Suffix (RLCP); Hamming distance. Starting from each of these dissimilarity matrices, a cluster analysis is conducted using the Ward algorithm. Then I derive a measure of agreement in classification between the LCA solution and the cluster solutions derived by the SA approaches. The agreement in classification is measured with the Rand index (Rand, 1971) that measures the proportion of couples of observations classified in the same group by two cluster solutions. The corrected version of Rand Index (Morey and Agresti, 1984) accounts for the agreement due to chance. Results are presented in table 2.1. A detailed description of the clustering method and the classification index is presented in section 2.6.3.

Table 2.1.: Agreement in classification between LCA and SA techniques

	Rand index	Corrected Rand index
OM with empirical costs	0.88	0.59
Longest common subsequence (LCS)	0.87	0.55
OM with constant costs	0.86	0.52
Dynamic Hamming distance (DHD)	0.86	0.50
Longest Common Prefix (LCP)	0.77	0.26
Longest Common Suffix (RLCP)	0.71	0.19
Hamming distance	0.71	0.19

In this example, Optimal Matching with empirical-derived costs gives the closest solution to the classes identified by LCA. The rand index is 0.88 meaning that among all the possible pairs of observations, almost the 90% are classified in the same group using the two methods. The corrected version of the Rand index accounts for the proportion of agreement due to chance and reduces the percentage of couples classified in agreement to 59%. The LCS distance does not imply any cost settings. The cluster solution obtained with this method is very similar to the OM solution (0.87 Rand index, 0.55 the corrected version). Using constant costs does not substantially decrease the agreement with respect to the OM version with empirical costs. Also the use of dynamic costs based on the age of the respondent does not change the percentage of agreement between the two classification. On the other hand, the cluster solutions obtained with the remain-

ing distances (LCP; RLCP and Hamming distance) diverges substantially from the LCA solution presented in the paper by Amato et al. (2008).

This example does not motivate the use of a particular distance respect to the others, but gives a first indication on the consistence of different statistical methods for life course analysis. In particular it is interesting to notice that, in this case, the two methods that lead to a closer solution to LCA are OM with transition costs and LCS. In the simulations presented in this chapter, I compare LCA with these two methods for sequence analysis. Although the different approaches for life course classification seem to be consistent (in particular between LCA and OM), it is not possible to draw any conclusion on the reliability of the methods if the generating mechanism of life course sequences is unknown.

2.6. A simulation study

I propose a simulation approach to study the factors affecting the goodness of LCA and SA techniques. The simulation procedure can be summarized in 4 steps:

1. Define typical groups of life course trajectories
2. Introduce variability in timing, quantum and sequencing
3. Classify individuals of the artificial dataset using Latent Class and Optimal Matching techniques
4. Compare classification obtained with the two techniques with the real groups

A simulation approach to test the reliability of SA techniques has been previously proposed by Wilson (2006) to test the performances of the *ClustalG* multiple alignment package. The simulation study proposed in this chapter, however, follows a different approach. Instead of starting from a stochastic generating mechanism, the reliability of SA techniques is tested increasing the level of heterogeneity among groups of sequences.

2.6.1. Defining typical groups of life course trajectories

Let us define 4 different groups of life course trajectories using a simple state-space composed by the states S,C,M. For each sequence, I set the length equal to 30 and S

as initial state. Then, I repeat every typical sequence 250 times obtaining an artificial dataset of 1000 observations. The dataset can be considered as a monthly (quarterly) collection of data indicating the marital/union status of an individual. One, for example, can consider S as single, C as cohabiting and M as married. Let us define 4 "typical" groups of sequences:

1. (S,10)-(C,10)-(M,10)
2. (S,20)-(C,5)-(M,5)
3. (S,10)-(C,5)-(M,10)-(C,5)
4. (S,20)-(M,10)

Where (X, t) indicates t periods in state X. Individuals from group 1 are single for 10 periods than they cohabit for 10 periods and then they stay in marriage until the end of the sequence. Groups differ for timing, quantum and order. For example, group 1 differs from group 2 because individuals exit state S earlier, from group 3 because of the order of states M and C and from group 4 because they experiences state C.

2.6.2. Introducing variability in the typical sequences

To test the reliability of the two methods, I introduce random perturbations in timing, quantum and sequencing of trajectories. The idea is to confound the latent groups modifying sequences with different sources of noise. Thus, I introduce a series of sequence operators that modify life trajectories. These operations introduce variability in the groups. Even if these operations have not a specific meaning in the social sciences, I tried to mimic some behaviors observed by individuals during the life course.

Let us define the following operators:

- Postponement

  With probability p (postponement rate), copy status from time t to time $t+1$

  S S S S S S S S S S **C** C C C C C C C C C **M** M M M M M M M M M M
  S S S S S S S S S S **S** C C C C C C C C C C **C** M M M M M M M M M M

- Slicing

  With probability p (slicing rate), exchange two subsequence of the same length

  S S S S S S S S S S **C C C C C** C C C C **C** **MMMM**M M M M M M
  S S S S S S S S S S **C** **MMMM**C C C C **C C C C C** M M M M M M

- Inversion

  With probability p (inversion rate), exchange all the elements C with elements M

  S S S S S S S S S S **C C C C C C C C C C** M M M M M M M M M M
  S S S S S S S S S S **MMMMMMMMMM**C C C C C C C C C C

- Mutation

  With probability p (mutation rate), substitute sequences status at time t with a random element of the alphabet.

  S S S S **S** S S S S S **C** C C C C C **C** C C C **M**M M M **M**M M M M M
  S S S S **M**S S S S S S **S** C C C C C **C** C C C **C** M M M **C** M M M M M

- Truncation

  With probability p cut sequence at time t, with t randomly chosen.

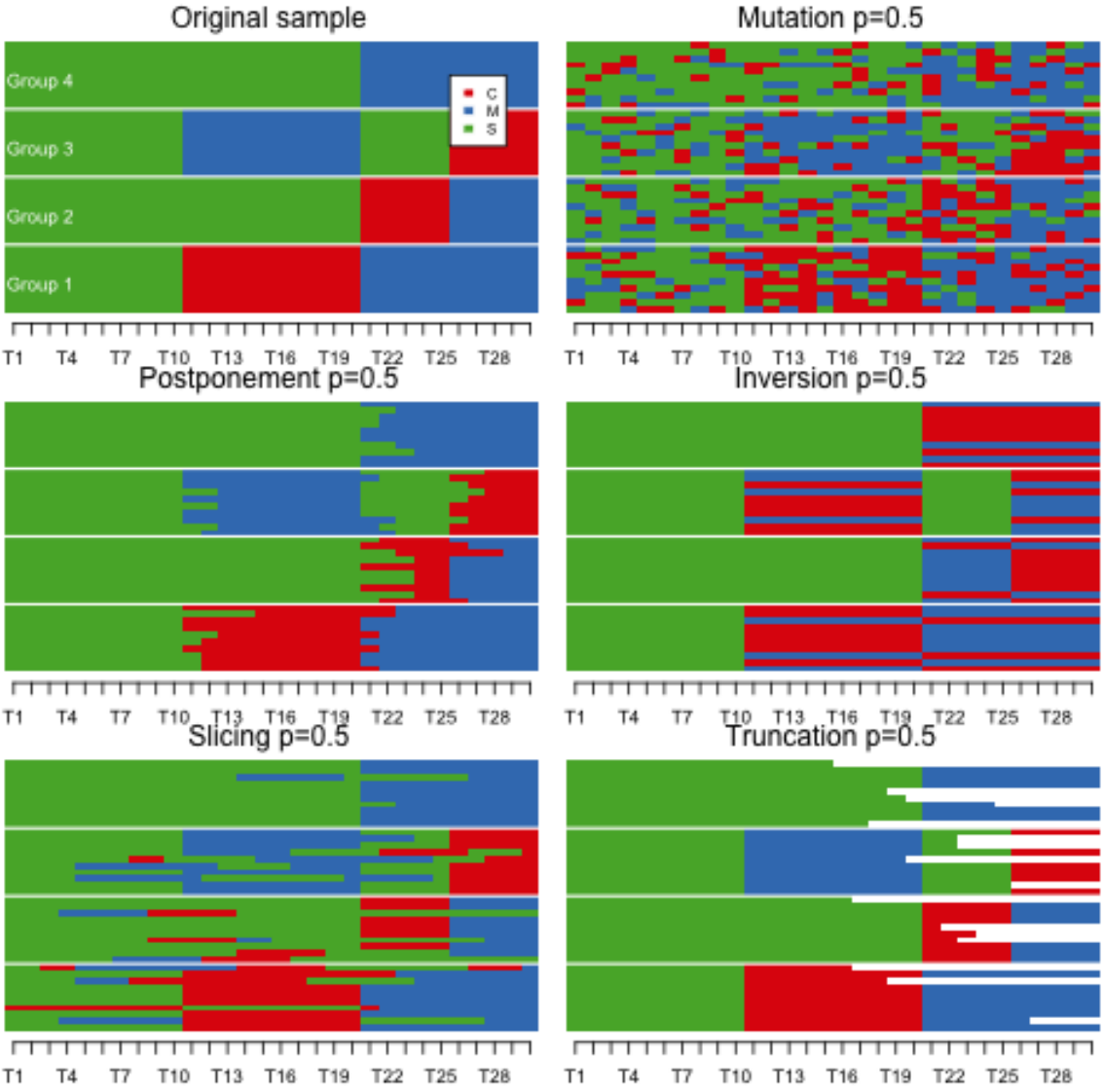
  S S S S S S S S S S C C C C C C C C C C M M M M **M**M M M M M
  S S S S M S S S S S S C C C C C C C C C C M M M

The operators proposed are meant to introduce variations in the different components of life course introducing variability among sequences. The general idea is to modify the sequences mimicking the behavior of real life course trajectories. For example, some individuals may postpone (or anticipate) a transition, while others invert the "order" in which events happen. Mutation does not have a direct life course interpretation, but it can be described as a source of measurement error, since it may occur that individuals are randomly misclassified across time. Using this disturbance strategy allows to test the reliability of different methods without assuming any generating mechanism of the data.

How to measure variability in Timing, Quantum and Sequencing

- **Timing** The *tempo* dimension of a transition is the timing in which a change of state occurs. The exit time from the first time is a crucial transition in many demographic studies (i.e. leaving parental home, entering the first union, having the first child). As a naive indicator of timing, I define the age at first transition. The standardized indicator τ expresses the proportion of a life sequence spent in

Figure 2.3.: Effects of different sequence operators



the initial status. Precocious individuals have a low value of τ, on the contrary τ increases with postponement.

$$t_{min} = min\{s_{(t-1)} \neq s_t\} \qquad t = 1, \ldots, T$$

$$\tau = t_{min}/T$$

- **Quantum**. The number of events is a key element that characterizes a life course trajectory. The concept of *Quantum* indicates the likelihood of an individual to experience transitions. A simple indicator can be expressed by the overall number of transitions. The standardized value ρ indicates the number of transitions per time period.

$$\rho = \frac{\#\{s_{(t-1)} \neq s_t\}}{T}$$

- **Sequencing** The order in which events occur is crucial in the study of life course. For example, it may be relevant to study the divergence of a life trajectory from the normative course of transition. For this reason, I propose as an indicator, the number of non-normative transition. That is, the transitions that diverge from a given sequence of events considered normative in the society. The standardized value ς indicates the proportion of normative transitions over the total number of transition.

$$\varsigma = \frac{\text{Number of normative transitions}}{\text{Total number of transitions}}$$

The three indicators range between 0 and 1.

These operators modify different dimension of life courses. Postponement introduces a major change in timing while the other two dimensions remain unaltered. Inversion modifies only the order of events because it transforms an entire category of events into another. Slicing modifies both the order and the quantum of events. Last, mutation has a massive effect on quantum, but it affects also the other two dimensions introducing completely random variations. The effect of the sequence operators are illustrated in figure 2.4.

2.6.3. Classification

Once defined a new dataset, modified by the previous "sequence operators", it is possible to apply the alternative classification procedures. While LCA requires less specifications by the researcher, in sequence analysis one need to specify the costs (only in case of OM) and the clustering procedure. Following the most common approach in SA for demographic studies, I estimate OM distances using costs proportional to transition rates

Figure 2.4.: Effects in timing, quantum and sequencing. Mutation, Postponement

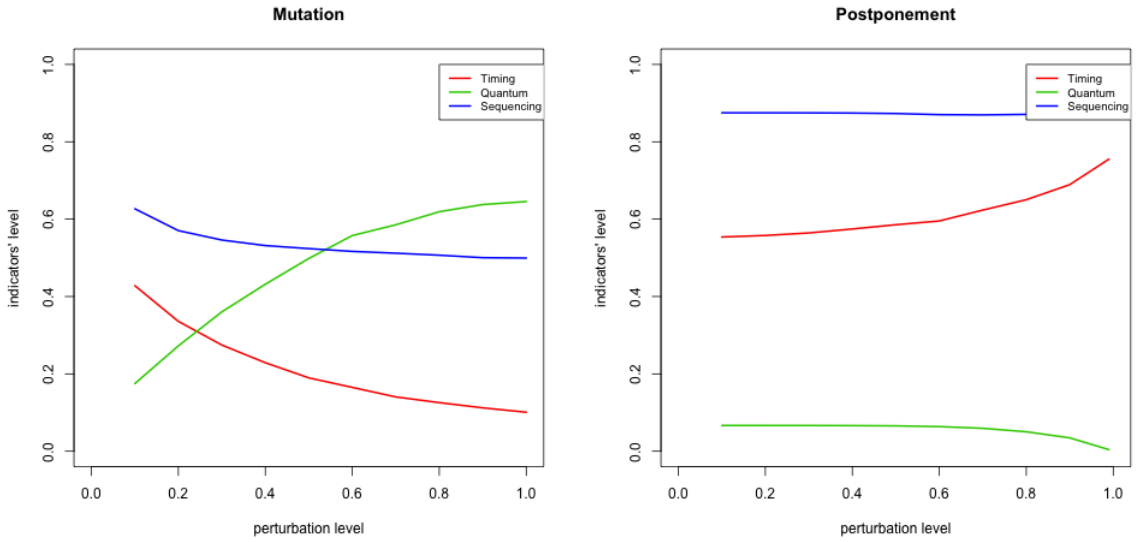
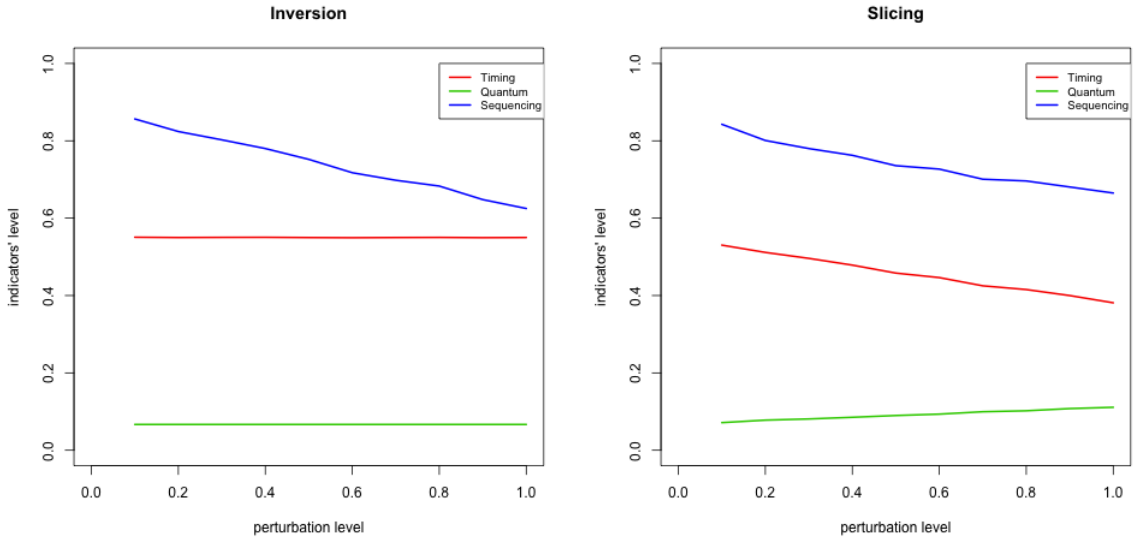


Figure 2.5.: Effects in timing, quantum and sequencing. Inversion, Slicing



and I use standard Ward algorithm for clustering. Ward clustering algorithm (Ward, 1963) can be briefly described as it follows. Consider N individuals to be clustered according to their sequences. Let $d(i, j)$ denote the distance between the ith and the jth individual sequences. The total dispersion, i.e. the amount of dispersion within the whole

data set, is usually measured as $T = \sum_{i,j} d(i,j)$. Suppose now that the whole sample is partitioned into G clusters. The dispersion within the gth cluster is $W_g = \sum_{i,j \in g} d(i,j)$, and the dispersion within the G groups can be summarized as $W_G = \sum_{g=1}^{G} W_g$. The adequacy of a clustering solution is often evaluated by referring to $R_G^2 = 1 - W_G/T$, which is the proportion of the total dispersion accounted for by the G clusters. By construction, if $G-1$ clusters are obtained by joining two clusters, say g_L and g_R, out of a number of G, into a single one g, it follows that $W_G < W_{G-1}$, and $R_G^2 > R_{G-1}^2$. Hierarchical agglomerative clustering algorithms proceed by sequentially joining pairs of clusters: they differ in the criterion that is followed to select which clusters must be joined. In Wards algorithm the two clusters to be joined are selected by minimizing the increase in the within-groups dispersion consequent on the reduction of the partitions degree:

$$\Delta(g|g_L, g_R) = W_g - W_{gL} - W_{gR} = W_{G-1} - W_G \tag{2.2}$$

or, equivalently, by minimizing $\Delta R_{G-1}^2 = R_G^2 - R_{G-1}^2$. The result of this hierarchical procedure is a sequence of (nested) clusters solutions having a decreasing number of clusters, $\{P_{max}, P_{max-1}, ..., P_1\}$, max being the maximum number of clusters that we can define, coinciding with N, the total number of cases. Given a partition P_G, the P_{G-1} partition is determined by (conditionally) maximizing R_{G-1}^2, i.e. by minimizing the decrease in the R^2 due to the reduction of the number of clusters.

Latent Class has been conducted setting binary variables in each time period indicating if the individual is single (S), cohabiting (C) and married (M). To avoid local maxima I run the model 3 times and I choose the model with the minimum BIC. For practical purposes, both the number of classes and the number of clusters is set fixed. The analyses conducted varying the number of classes give similar results in terms of classification performances.

2.6.4. Classification performances

The goodness of classification is measured examining the association rate between the classes obtained by the two methods and the original groups. I measure how the association between the real and the actual groups changes according to different levels of

disturbance. Association rate is measured with a modified version of Rand index (Rand, 1971). Rand index measures the proportion of couples of observations that are classified in the same group by two (or more) judges. Suppose that in the population of interest, there are k_1 clusters in the first solution and k_2 clusters in the second. Let P_{ij} be the probability that a randomly selected individual is classified in cluster i in the first solution and cluster j in the second solution. Rand's statistic is defined to be the probability that a randomly selected pair is classified in agreement. This probability equals

$$P_s = \sum\sum P_{ij}^2 + \sum\sum P_{ij}(1 - P_{i+} - P_{+j} + P_{ij}) \tag{2.3}$$

$$= 1 - \sum P_{i+}^2 - \sum P_{+j}^2 + 2\sum\sum P_{ij}^2 \tag{2.4}$$

This measure of agreement has the advantage that can be used even if the size of the two clusters (k_1 and k_2) differ. On the other hand, Rand index makes no correction for chance agreement. Therefore, it is not possible to tell whether a specific value of P_s is "large" or "small", because its value when individuals are classified at random (i.e. $P_{ij} = P_{i+}P_{+j}$) is not zero, and depends on P_{i+} and P_{j+}. This can constitute a disadvantage when the replicability of different classifications are being compared. In this chapter, I use the corrected version of the Rand Index (Morey and Agresti, 1984) that properly takes into account the proportion of agreement due to chance. The corrected version of Rand's statistic equals

$$\Omega = \frac{2\sum\sum P_{ij}^2 - 2(\sum P_{i+}^2)(\sum P_{+j}^2)}{\sum P_{i+}^2 + \sum P_{+j}^2 - 2(\sum P_{i+}^2)(\sum P_{+j}^2)}. \tag{2.5}$$

This statistic equals one for perfect agreement, $\Omega = 0$ for chance agreement, and $\Omega < 0$ when agreement is less than expected by chance.

2.7. Simulation results

I simulated 1000 samples for each sequence operator applying different level of disturbance. For each sample, I estimate a latent class model with 4 classes and I calculated OM and LCS matrix of dissimilarity. Then I apply a cluster analysis using Ward algorithm to classify individuals in 4 groups. The groups obtained are compared with the original groups using the corrected Rand index. Figure 2.6 and table 2.3 report the average rate

of agreement between the original groups and the results obtained by latent class analysis and sequence analysis (OM and LCS). Results show that classification is sensitive to the transformations inducted by sequence operators. With the increasing of variability in the sample, the classification goodness decreases. As expected, the performances of all the methods decrease rapidly with random mutation. Mutation, in fact, can be considered a benchmark since it introduces the maximum amount of variation. The agreement rate under postponement decreases more slowly. In particular small postponement rates do not seem to affect the probability of good classification. However, precision decreases with higher disturbance levels. Postponement principally affects timing, since it extends the amount of time spent in the initial status. But a massive postponement has also an effect in quantum, since it reduces the amount of transitions in trajectories and reduces the variability between different groups of sequences. Inversion has the maximum confounding effect at rate 0.5. At that point, exactly half of sequences get all "C" inverted with "M" and vice-versa. With greater inversion rates, the order of sequences changes and, in turn, variability within groups is reduced. Therefore, classification becomes straightforward. Slicing has an effect both on sequencing and quantum of life course and the classification decreases almost linearly. The performances of classification under truncation follow a U-shape. An increase in truncation rate affects the number of censored individual sequences. It follows that high truncation rates are associated with sequences that are shorter in average. For this reason (since truncation is randomly assigned to the second half of the sequence), I observe an increase in classification agreement when the truncation rate is high.

The results obtained by our simulations suggest some considerations about the reliability of these classification methods. First, there are no evidence of a methodology that have superior performances under all the sources of variation. In fact we do not observe a methodology that perform better in all the cases. Despite that, according to our simulations, LCA has better performances under mutation and truncation. On the other hand, SA shows greater agreement in inversion and slicing. Results from postponement indicate a substantial equivalence of the techniques with slightly better results for sequence analyses. Second, the classifications with latent class analysis seem to be less

precise. Using simulated data it is possible to have an indication on the variability of the estimated agreement rates. Under all the sources of error, the results obtained with LCA exhibit more variability. Third, the differences between OM and LCS are minimal. Both the methods, in fact, produce very similar results. Although the two distances are qualitatively different, the results obtained in all the sources of variability are very similar.

To summarize the results I propose a measure of the overall performance. Let R be the number of simulations, and $\Omega_r^{\{LCA;OM;LCS\}}$ the corrected Rand index for the sample r under different sequence operators. A simple index of the overall goodness of classification is the expected Rand index $\bar{\Omega}$.

$$\bar{\Omega} = \frac{1}{R}\sum_{r=1}^{R}\Omega_i \tag{2.6}$$

$\bar{\Omega}$ can be interpreted as the expected agreement between the true groups and the estimated classification. Table 2.2 summarizes the results. Sequence analysis techniques seem to have better performances under postponement, inversion and slicing. Latent class analysis gives better results under mutation and in case of data truncation.

Table 2.2.: Classification rate

	Mutation	Postponement	Inversion	Slicing	Truncation
LCA	0.586	0.713	0.566	0.427	0.608
OM	0.520	0.735	0.638	0.632	0.549
LCS	0.509	0.737	0.647	0.646	0.552

Figure 2.6.: Classification results

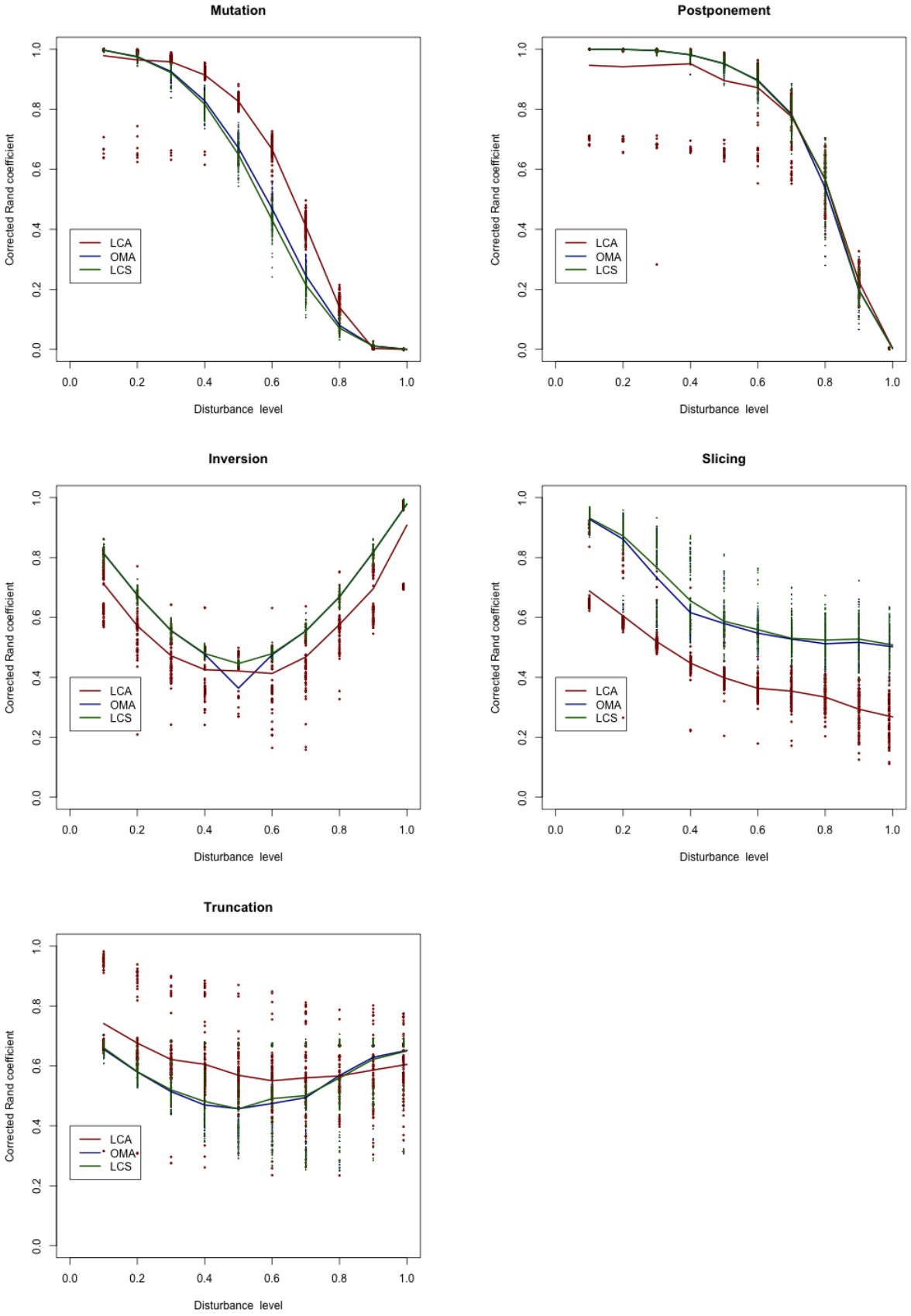


Table 2.3.: Simulation results

		Inversion			*Postponement*			*Mutation*			*Slicing*			*Truncation*		
		LCA	OM	LCS	LCA	OM	LCS	LCA	OM	LCS	LCA	OM	LCS	LCA	OM	LCS
0.1	Mean	0.7100	0.8120	0.8120	0.9470	1.0000	1.0000	0.9790	0.9970	0.9970	0.6870	0.9270	0.9320	0.7400	0.6550	0.6590
	Var	0.0077	0.0005	0.0005	0.0131	0.0000	0.0000	0.0064	0.0000	0.0000	0.0087	0.0005	0.0005	0.0176	0.0003	0.0023
0.2	Mean	0.5700	0.6730	0.6730	0.9420	1.0000	1.0000	0.9650	0.9750	0.9730	0.6040	0.8590	0.8710	0.6750	0.5790	0.5800
	Var	0.0063	0.0004	0.0004	0.0143	0.0000	0.0000	0.0078	0.0001	0.0001	0.0062	0.0068	0.0049	0.0146	0.0009	0.0008
0.3	Mean	0.4720	0.5560	0.5560	0.9470	0.9950	0.9950	0.9580	0.9250	0.9210	0.5190	0.7310	0.7670	0.6220	0.5130	0.5190
	Var	0.0047	0.0003	0.0003	0.0163	0.0000	0.0000	0.0041	0.0005	0.0005	0.0018	0.0146	0.0134	0.0102	0.0021	0.0024
0.4	Mean	0.4250	0.4760	0.4790	0.9520	0.9810	0.9820	0.9140	0.8270	0.8150	0.4470	0.6150	0.6540	0.6040	0.4690	0.4800
	Var	0.0061	0.0007	0.0002	0.0111	0.0002	0.0002	0.0026	0.0011	0.0011	0.0027	0.0064	0.0107	0.0137	0.0053	0.0067
0.5	Mean	0.4220	0.3640	0.4460	0.8930	0.9500	0.9520	0.8250	0.6690	0.6480	0.3980	0.5790	0.5870	0.5690	0.4560	0.4550
	Var	0.0043	0.0028	0.0000	0.0187	0.0006	0.0005	0.0013	0.0021	0.0016	0.0023	0.0037	0.0038	0.0071	0.0113	0.0115
0.6	Mean	0.4130	0.4760	0.4790	0.8720	0.8960	0.8940	0.6630	0.4660	0.4290	0.3630	0.5470	0.5590	0.5500	0.4750	0.4930
	Var	0.0075	0.0006	0.0002	0.0123	0.0015	0.0016	0.0025	0.0025	0.0025	0.0018	0.0027	0.0033	0.0111	0.0142	0.0151
0.7	Mean	0.4700	0.5560	0.5560	0.7720	0.7820	0.7760	0.4080	0.2440	0.2130	0.3540	0.5270	0.5300	0.5610	0.4940	0.5000
	Var	0.0075	0.0005	0.0005	0.0085	0.0025	0.0042	0.0020	0.0016	0.0017	0.0026	0.0025	0.0026	0.0140	0.0176	0.0194
0.8	Mean	0.5790	0.6700	0.6700	0.5630	0.5350	0.5620	0.1390	0.0790	0.0700	0.3330	0.5130	0.5240	0.5670	0.5690	0.5600
	Var	0.0057	0.0006	0.0006	0.0065	0.0079	0.0066	0.0015	0.0003	0.0004	0.0024	0.0024	0.0033	0.0089	0.0142	0.0164
0.9	Mean	0.6980	0.8190	0.8190	0.2240	0.1950	0.1950	0.0030	0.0110	0.0110	0.2940	0.5170	0.5280	0.5870	0.6290	0.6230
	Var	0.0085	0.0005	0.0005	0.0020	0.0018	0.0019	0.0000	0.0000	0.0000	0.0039	0.0041	0.0040	0.0104	0.0076	0.0089
0.99	Mean	0.9080	0.9790	0.9790	0.0030	0.0030	0.0030	0.0000	0.0000	0.0000	0.2680	0.5020	0.5080	0.6050	0.6520	0.6500
	Var	0.0142	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0041	0.0039	0.0037	0.0086	0.0048	0.0051

2.8. Discussion

In the last decade, holistic methods for life course analysis have become more and more common. Instead of focusing only on life course transitions, the object of the study is the entire life trajectory. Life course trajectories can be described as categorical time series where time is associated to life states. Using longitudinal or retrospective data, it is, in fact, possible to describe individuals' life courses as age-referenced sequences of events. Rather than modeling directly the probability of the occurrence of a particular event, holistic methods attempt to individuate important patterns in the data using a data mining approach. In the literature of life course analysis we can distinguish two principal approaches: latent class analysis and sequence analysis.

It is not clear, however, how reliable are these methods in detecting effectively patterns in the data. A bigger critic that have been moved to these techniques is exactly their reliability and the difficulties in testing it. For this reason, I propose a simulation approach to investigate the reliability of classification techniques in life course analysis. Furthermore, I propose a method to simulate life sequencing without making any assumptions on the generating mechanism of the data. Starting from homogeneous groups of life trajectories, I introduce different sources of variability that, mimicking individuals' behavior, transform life courses in different dimensions. This approach allows to test if there are substantial differences in detecting groups of life trajectories.

Our simulation results show that the two methods are consistent. Although I do not found the absolute superiority of a method respect to the other, our results show that OM and LCS seem to have better performances when life course sequences are modified in the ordering of transitions (inversion and slicing). On the other hand, LCA has better results when the variations are completely random (mutation). Although random mutation may be common in some scientific fields, i.e. biology or information theory, a random disturbance appears to be quite unlikely in life course analysis. Individuals may experience unexpected events in life course, but usually these events are associated with a duration and rarely have no effect on the following part of the life trajectory. Nevertheless, mutation can be interpreted as a measurement error, since individuals may be

misclassified during repeated measurements.

Overall, the results obtained in this chapter justify the use of sequence analysis (in particular OM and LCS) for the study of life course. Our sequence operators do not cover all the possible variation that can occur in life course. That otherwise would be impossible. Also, life course classification may be influenced by other factors (i.e. the length of sequences, the dimension of the state-space and the classification algorithm). Despite that, this study presents some limitations, it represents one of the first attempt to test the reliability of holistic methods for life course analysis.

3. What explains the heterogeneity in early family trajectories? A non-parametric approach for sequence analysis

3.1. Introduction

During the last decades, there have been profound changes in partnering and childbearing in the United States, including changes in cohabitation and non-marital fertility. Women in their early 20s have been particularly affected (Schoen et al., 2007; Amato et al., 2008); as a result, it is important to examine divergences in the initial years of early adulthood. In terms of family transitions, those years are very "dense" (Rindfuss et al., 1987), with more demographic events occurring than during any other part of the life course. The under-25 age group exhibits great heterogeneity in family formation behavior, with some women postponing all family-related transitions, others making commitments (e.g., cohabitation), and still others making choices with enduring consequences (e.g., becoming a parent). Life course trajectories are in part the outcome of individual life planning, sometimes with the participation and help of the parents and/or partner, but they are also influenced by the social origin and the dynamic context around young adults.

It is not completely clear, however, how much of the heterogeneity observed in life courses can be explained by demographic and socioeconomic variables. It is reasonable to expect that individuals who share similar characteristics before starting family formation experiences will have similar family trajectories. On the contrary, greater variability should be expected by individuals from different sociodemographic backgrounds. In this chapter, I attempt to investigate the heterogeneity in family trajectories during early

adulthood using sequence analysis techniques. Sequence analysis gives a representation of the occurrence, the timing and the ordering of a set of events observed across time. Using a rich longitudinal dataset (Add-Health), I compute the monthly family trajectories of a sample of young women living in the United States. Examination of the simultaneous distribution of cohabitation, marriage and parenthood provides data for the trajectories. In each month, individuals can be classified as: Single (S), Single Parent (SP), Cohabiting (C), Cohabiting parent (CP), Married (M) or Married parent (MP). Using Optimal Matching Algorithm (OMA) (Abbott, 1995), it is possible to deliver a measure of dissimilarity between family formation trajectories. In general, to calculate a pairwise distance between two sequences, the number of minimum transformations (insertion, deletion, and substitution) necessary to transform one sequence into the other is tallied, each transformation is assigned a cost, and these costs are summed. The cost of a single substitution is derived empirically by data. The result of OMA is a matrix of pairwise dissimilarities that usually is the starting point for data reduction techniques, mainly clustering or multidimensional scaling (Piccarreta and Billari, 2007; Aassve et al., 2007)[1]. In this chapter, I propose a non-parametric procedure to test differences between groups of life trajectories. Without using data reduction techniques, I propose to study pairwise distances between observations. As standard ANOVA for linear models, I decompose the dissimilarity between trajectories in a component explained by a "model" and in a "residual" component. Then it is possible to evaluate a pseudo F-statistic and a pseudo-R^2 in order to evaluate the explanatory power of each variable. Suitable statistical tests are conducted following a non-parametric approach, approximating the distribution of the statistic under the null hypothesis with a permutation procedure.

The aim of this chapter is both substantive and methodological. I attempt to describe the impact of demographic and socioeconomic variables on the variability of life-course trajectories answering questions like: How similar are the life courses of individuals that share the same characteristics at the beginning of the transition? What are the variables that most influence the divergence of trajectories? Conversely, I propose a semi parametric procedure to model the variability of sequence analysis without using data reduction

[1]For a more exhaustive description of OMA, I refer to section 2.4

techniques.

3.2. Motivation and research questions

The variability of life trajectories is a central issue in life course sociology and demography. A large body of literature, in fact, is dedicated to analyzing the historical variations in life course between cohorts. Early adulthood and interactions with social context garner particular attention (see e.g. Hogan and Astone, 1986; George, 1993; Rindfuss et al., 1987). The majority of these studies analyze the evolution across time of the occurrence, timing and order of different markers of transition to adulthood (e.g. marriage, childbearing, leaving parental home, transition from school to work).

Most of the scholars argue that family trajectories in United States achieved a high level of uniformity by the 1960s. The modernization of society led to an increase in standardization of life courses throughout the 19th and 20th century. The continuous evolution in the organization of public services (in particular public educational systems) increased the age structuring of events and rendered the life course more orderly and predictable (Modell et al., 1976). It follows that, until the 1960s, a large majority of individuals experienced an identified set of ordered and age-graded family stages with very few of them getting out of sequence or delaying transitions.

The evolution of the variability in life course over the last decades is more controversial (Macmillan, 2005). Several studies suggest that life course became more variable and less uniform starting from the 60s. This process is often described as a "new individualization" in life course (Shanahan, 2000). As in the previous decades, individuals are still subject to institutional constrains but at the same time they are less tied to familiar and local contexts. In other words, life course has increasingly become the result of a deliberate plan. In particular, family formation has become, in many countries, a more extended sequence of events because some demographic phenomena (cohabitation, staying childless, living single, extra-marital parenthood) have become more accepted and practiced (Cherlin, 2004, 2005). Brückner and Mayer (2005) tested the de-standardization hypothesis, examining different German cohorts and concluded that most of the changes

in early life course are due to variation in family formation while education and labor force participation increase in homogeneity.

Sequence analysis can be very effective to quantify and analyze heterogeneity in life course. This methodology, in fact, permits one to analyze simultaneously variation in timing, occurrence and ordering of life course events. Sequence analysis techniques have been recently used in various studies to compare variability in life course between different countries and different cohorts. Elzinga and Liefbroer (2007), for instance, use the Fertility and Family Survey data on 19 countries to test the de-standardization hypothesis of family trajectories. Their results show that in most of the countries, family trajectories of young adults have become less similar. The variability is analyzed using the turbulence index that takes into account variations in ordering and sequencing of life course events. In a similar fashion, Fussell et al. (2007) analyzes the transition to adulthood in Australia, Canada and the United States. An entropy measure for life sequences analyzes the variability in life course. Furthermore, a sequence analysis approach has been recently used by Widmer and Ritschard (2009) to test how cohort and sex affect de-standardization in life course.

Although variability in life course is a major topic in life course analysis, most of the works focus on variability *between* cohorts. Only a limited number of studies focuses on variability *within* the same cohort. Family trajectories are influenced by a large array of socio-economic characteristics such as race, parents' education, family composition, (see e.g. Schoen et al., 2009; Landale et al., 2010; Schoen et al., 2007), and geographic context (South, 2001; Evans et al., 1992; Teachman and Crowder, 2002; Turney and Harknett, 2010). It is not clear, however how much of the variation within individuals can be explained by other variables. Are individuals who share the same characteristics more similar in terms of life trajectories? How can we measure this association? It is reasonable, in fact, to expect a certain grade of homogeneity between individuals with similar traits. Jackson and Berkowitz (2005), for instance, examine the variation in occurrence and sequencing of work and family events and find a great similarity within same sex and race groups.

The aim of this chapter is to examine, in detail, the variability in family formation

trajectories within a cohort of young US women. I focus on two sources of variability: background characteristics and geographical context. The underlying idea is that individuals who share similar characteristics at the beginning of adulthood might have similar family formation trajectories. The objective of the chapter is, therefore, to test if heterogeneity in life course can be partially explained by these two sets of variables. In particular, I individuated four separate hypotheses that I propose to test. In the first two, I test the direct effect of background characteristics derived from the family of origin and the effect of the geographical context. In the third and fourth, I test the interaction between socio-economic resources and standardization.

Hypothesis 1: Background characteristics I hypothesize that individuals with similar backgrounds are more likely to experience similar family trajectories. In particular, I suppose that characteristics such as race/ethnicity, parents' education, parents' birthplace, family income and religiosity affect family patterns. I expect, therefore, that women who share one or more of these characteristics are more similar in terms of family trajectories. I also suppose that the influence of the family of origin is crucial in determining family trajectories. I expect therefore that young women living in the same household (sisters, half-sisters and twins) are more similar than couples of individuals randomly chosen from the sample.

Hypothesis 2: Geographical context I hypothesize that individuals living in similar geographical contexts during childhood and early adolescence are more likely to experience similar family trajectories. I examine the effect of geographical context looking at characteristics such as median income, poverty level, unemployment rate and percentage of foreign-born individuals. I expect that women who used to live in the same geographical area during childhood and adolescence will have more family trajectories that are similar in early adulthood. The geographic context may describe the social capital, which young women draw on during the transition to early adulthood. Also, I consider schools and their typology (public versus private). In the case of geographic context, I expect to see less variation between trajectories in smaller geographic areas. Using the geographical

information of the sample, I test if individuals are more similar in smaller geographical contexts. The geographical context components analyzed are: state, county, tract, and block. If the geographical context has an effect on sequences' variability, I expect that individuals are more homogeneous at block level with respect to greater geographical detail.

The two hypotheses above are tested comparing distances among pairs of individuals in the same group with distances of pairs randomly chosen from the sample. Using Optimal Matching distances it is possible to establish a measure of dissimilarity among life course sequences. In a linear model framework, it is common to compare categorical variables using models of analysis of variance (ANOVA). In this chapter, I present a procedure that extends ANOVA to sequence analysis. Using a non-parametric approach, it is possible to conduct statistical tests in order to verify if the observed divergences are statistically significant.

This procedure allows to verification that some categories exhibit a particular degree of homogeneity compared to others. It is possible, for example, to investigate characteristics that might be associated with greater standardization. Analyzing variability within the same cohort allows investigation of the precursors of de-standardization in life course. The analysis is focused on the interactions between socio-economic resources and standardization.

Hypothesis 3: Social class increases de-standardization I hypothesize that women with higher socio-economic status experience more de-standardized family trajectories. I expect that women with more educated and wealthy parents or those who used to live in higher class neighborhoods experience more heterogeneous family trajectories. Young women from higher social class can draw on more resources, in terms of human and social capital, during the transition to early adulthood. This could lead to a more individualized life course since life course may be less influenced by socio-economic constraints. On the contrary, life course is more likely to be determined only by individual preferences. Under

this hypothesis, I expect that women from higher social class are less likely to adhere to traditional, age-normed family rules. I expect, therefore, that women from higher social class are more likely to experience less standardized family patterns. On the contrary, women with less resources may be constrained to follow a more standardized life course.

Hypothesis 4: Social class increases standardization This last hypothesis represents the complementary version of hypothesis 3. I hypothesize that social class does not increase de-standardization, but on the contrary, increases homogeneity and standardization in terms of family trajectories during early adulthood. Therefore, I expect that young women from higher social classes experience more homogeneous life trajectories. Social class, indeed, is associated with a more institutionalized life course; in particular, social class increases the time spent in education. This is generally associated with a delay in family formation and a more age-normed family trajectory. Young women with more socio-economic resources are less likely to drop-out of school or to experience early family unions and early childbearing. Under this hypothesis, family background and social capital increase standardization and predictability in life course. On the other hand, women from lower social classes can draw on less socio-economic resources. This can prevent them from achieving the desired pattern of family formation. Under this hypothesis, social class protects from unplanned events during early adulthood and increases homogeneity in family trajectories. For these reasons, I expect to observe less heterogeneity among women with educated parents, greater family income and living in better neighborhood.

Analyzing the variability in family formation *within* a cohort of young women is relevant for a number of different reasons. First, it may help to quantify how much of the variability in life course observed in the last decades is imputable to variations occurring among subgroups of populations. It is possible, in fact, that the process of de-standardization takes place at different paces in different groups of populations as determined by race or education levels for example. Second, it may help to shed a light on

the link between social stratification and life course. It may help, in fact, to understand how socio-economic status is connected to standardization in life course. Furthermore, studying the entire trajectory can help to study simultaneously heterogeneity in timing quantum and sequencing of family transitions.

3.3. Data

3.3.1. Sample

The National Longitudinal Study of Adolescent Health (Add Health) is a school-based, nationally representative sample of U.S. students in grades 7 through 12 in 1994. Nearly all the respondents were born in the years 1976 through 1982. The Add Health data include four waves of in-home interviews, which were conducted in 1995 (Wave I), 1996 (Wave II), 2001-2002 (Wave III) and 2008-2009 (Wave IV). The data for the present study are taken from Waves I and IV. Of the 10,480 women interviewed in 1995, 8,015 were also interviewed during Wave IV. Since we restrict our analysis to women that we can observe from age 18 to 26, the final sample size becomes 6,974. Using retrospective questions from wave IV, I reconstructed the family biographies of women from age 15 to 26.

The Add-health study includes a sample of pairs selected in the same household. The pairs interviewed at wave I is 2,553. The sample is composed by full-siblings, half-siblings, twins (monozygotic MZ and dizygotic DZ), and non-related pairs (mostly cousins). In the analysis presented in this chapter, the number of individuals with at least another member of the household included in the sample is 1,956.

Incorporating systematic sampling methods and implicit stratification into the Add Health study design ensured that this sample is representative of US schools with respect to region of country, urbanicity, school size, school type, and ethnicity. In addition, data are geo-referenced and can be associated with different geographical levels (states, counties, census tracts and census blocks).

3.3.2. Variables

Family trajectories

Life course trajectories are represented by monthly combinations of union and childbearing states from age 15 to age 26. I designed the state space to take six possible values: Single (S); Single Parent (SP); Cohabiting (C); Cohabiting Parent (CP); Married (M) and Married Parent (MP). The months in which family events take place are defined using retrospective questions. In sequence analysis, each life-course or trajectory is represented as a string of characters (also numerical), similar to the one used to code DNA molecules in the biological sciences. I calculated pairwise distances using Optimal Matching Algorithm. Substitution costs are constant and imputed using the inverse of transition probability. The resulting output of the procedure is a dissimilarity matrix. The matrix is symmetric and has value 0 in the diagonal[2].

Background characteristics

Background characteristics of the respondents were collected in Wave I. Variables are categorical.

- *Race/ethnicity* (White, Hispanic, Black, Asian)
- *Family structure at Wave I* (living with both biological parents, living with a step parent, living with a single parent, other type of family)
- *Parents' education* (completed college, some college, high school, less than high school, unknown)
- *Parents' birthplace*(both parents born in US; at least one immigrant parent)
- *Religiosity at Wave I* (attended church or any religious ceremony at least once a week; less than once a week)

[2]For a complete description of costs definition see section 2.4.1

Geographical variables

Add-health data include contextual information for different geographical levels. For most respondents participating in the Add Health in-home survey, Wave I and Wave II home locations were identified. When possible, these locations have been geocoded in order to link them to their block group census areas[3] and their census tracts[4]. The availability of block group level data in the 1990 Census of Population and Housing for each of these areas has allowed the creation of a contextual data files corresponding to the two waves of data collection in the Add Health in-home survey. For the analysis of this chapter, I consider some economic indicators at block level detail. Variables have been then categorized in three categories according to their tercile. In addition to block-level variables, the typology of school (public or private) is taken in consideration.

- *Unemployment rate* (high; medium; low)
- *Median income* (high; medium; low)
- *Poverty level* (high; medium; low)
- *Percentage immigrants* (high; medium; low)
- *School type* (public; private)

3.4. Analysis of variance for life course sequences

I present a method to evaluate the association between a dissimilarity matrix and a set of categorical variables. This method has been introduced in ecology by Anderson (2001b) and McArdle and Anderson (2001) to analyze ecosystems. A permutation ANOVA has also been used by Zapala and Schork (2006) to evaluate the similarity between pairs

[3]The block group is a U.S. Bureau of the Census defined geographic area, which in 1990, averaged 452 housing units, or 1,100 people. It is the lowest level of geography for which the Census Bureau publishes sample data, and thus captures the most localized available contextual characteristics of the areas where individuals live.

[4]A census tract is a small locally defined statistical area within selected counties, generally having stable boundaries and, when first established by local committees, designed to have relatively homogeneous demographic characteristics. Census tracts do not cross county boundaries. They are generally defined for metropolitan areas and other highly populated counties and usually contain between 2,500 and 8,000 people.

of individual samples using high-dimensional genomic data. Furthermore, Studer et al. (2010) apply the same method to sequence analysis in social sciences.

The method presented is a generalization of the analysis of variance (ANOVA) in the case of metric and semi-metric measures. As in standard ANOVA models, the objective of the analysis is to partition the observed variance into components based on different sources of variation.

3.4.1. The univariate case

The basic idea of analysis of variance is to compare variability within groups versus variability between different groups, using the ratio of the F-statistic. The larger the value of F, the more likely it is that the null hypothesis (H_0) of no differences among the group means is false. For univariate ANOVA, partitioning of the total sum of squares, SS_{tot}, is achieved by calculating sums of squared differences between individuals and their group mean (SS_W, the within-group sum of squares) and between group means and the overall sample mean (SS_B, the within-group sum of squares).

$$SS_{tot} = SS_B + SS_W \tag{3.1}$$

In standard analysis of variance, the assumption is that observations are drawn from a (multivariate) normal distribution. In this case, the relation between sum of squares and euclidean distances is straightforward. The total deviance (SS_{tot}) can be expressed as the sum of pairwise euclidean distances de_{ij}.

$$SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - y_j)^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n}(d_{ij}^e)^2 \tag{3.2}$$

The output of a sequence analysis is generally a dissimilarity matrix. Anderson (2001b) and McArdle and Anderson (2001) propose to substitute d_{ij}^e with a generic distance (metric and semi-metric) in order to obtain a pseudo F-statistic.

Let D be a matrix of dissimilarities with elements d_{ij}, where d_{ij} represents the measure of dissimilarity between individuals i and j. Consider a simple case when the groups are composed by the same number of observations n. It follows that $N = an$ is the total

number of observation, and a is the number of groups. The total sum of squares is

$$SS_{tot} = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2. \tag{3.3}$$

In a similar fashion, the within-group or residual sum of squares is

$$SS_W = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2\delta_{ij} \tag{3.4}$$

where δ_{ij} takes value 1 if observation i and observation j are in the same group; otherwise it takes a value of zero. Then $SS_B = SS_{tot} - SS_W$ and a suitable pseudo F-statistic is

$$F = \frac{SS_B/(a-1)}{SS_W/(N-a)} \tag{3.5}$$

In an analog way is possible to calculate the amount of variability explained by the model with a pseudo R^2 equal to

$$R^2 = \frac{SS_B}{SS_{tot}}. \tag{3.6}$$

3.4.2. The multivariate case

Consider the multivariate case where p variables are measured, simultaneously, for each of n replicates in each of the a groups. Let $X_{(N\times p)}$ be a matrix of explanatory variables with p parameters associated to the N observations.

Traditional multivariate analysis of variance (MANOVA) assumes that observations $Y_{(N\times p)}$ are drawn from a multivariate normal distribution and the associated distance is the euclidean measure.

McArdle and Anderson (2001) show that the standard MANOVA can be extended in the case of non-euclidean distances. Let 1 be an array of length N with every elements equal to 1 and A matrix such that $a_{ij} = -\frac{1}{2}d_{ij}^2$. The sum of squares SS_{tot} can be written as the trace of G, where G is the Gower's centered matrix (Gower and Krzanowski, 1999).

$$G = (I - \frac{1}{N}11')A(I - \frac{1}{N}11') \tag{3.7}$$

As in the univariate case it is possible to partition the sum of squares into a component explained by the model and a residual one.

McArdle and Anderson (2001) show that the two quantities can be written as indicated in eq.3.8 and eq.3.9, with $H = X(X'X)^{-1}X'$ known as the hat matrix in the linear regression model.

$$SS_B = tr(HGH) \tag{3.8}$$

$$SS_W = tr[(I - H)G(I - H)] \tag{3.9}$$

Analogously to univariate ANOVA it is possible to calculate a pseudo F statistic.

$$F = \frac{tr(HGH)/(p-1)}{tr[(I-H)G(I-H)]/(N-p)} \tag{3.10}$$

If D is a matrix of euclidean distances, then $G = (YY')$ and F is equivalent to F in the standard MANOVA setting.

More generally, it is possible to compare two nested models in order to test the contribution of the inclusion (exclusion) of one parameter to the model.

Let indicate with c the complete model with p variables that is compared to the reduced model r composed by $r < p$ variables. Then, the pseudo F statistic is equal to

$$F_r^= \frac{SS_{B_c} - SS_{B_r}/(p-r)}{SS_{W_c/(n-p-1)}}. \tag{3.11}$$

The expression in eq.3.11 may be used to select the model using a *backwars* or *forward* selection procedure.

Assessing the statistical significance of the F-statistic

Since this procedure can be used with any metric or semi-metric distance, we can calculate the pseudo F-statistic and R^2 starting from a matrix of pairwise Optimal Matching distances. However, once F is calculated, we need a statistical procedure to test if the observed differences between groups are significantly different.

Since Optimal Matching (or other sequence analysis distances) differ substantially in distribution from euclidean distances, it is not possible to assume that F follows a Fisher distribution as in the standard linear model. For this reason, the statistical significance

of the F-statistic is evaluated using a permutation approach.

Suppose the null hypothesis is true and the groups do not differ substantially (i.e. life trajectories are very similar). If this were the case, then our life course sequences (rows) would be exchangeable among the different groups. Thus, the labels on the rows that identify them as belonging to a particular group could be randomly shuffled (permuted) and a new value of F obtained (called, say, F^{π}). This random shuffling and recalculation of F^{π} is then repeated for all possible re-orderings of the rows relative to the labels. This gives the entire distribution of the pseudo F-statistic under a true null hypothesis for our particular data. Comparing the value of F obtained with the original ordering of the rows to the distribution created for a true null by permuting the labels, a P-value is calculated as

$$P = \frac{(\text{No. of } F^{\pi} \geq F)}{(\text{No. of } F^{\pi})} \tag{3.12}$$

The original observed value of F is then a member of the distribution of F^{π} under permutation (i.e. it is one of the possible orderings). Usually, *a priori* significance level of $\alpha = 0.05$ is used for interpreting the significance of the results, as in other statistical tests. It is also possible to view the P-value as a measure of confidence concerning the null hypothesis (Fisher, 1955; Freedman and Lane, 1983). With a groups and n replicates per group, the number of distinct possible outcomes for the F-statistic in a one-way test is $(an)!/(a!(n!)^a)$ (Clarke, 1993). Usually p is calculated using a large random subset of all possible permutations since it is not practical to calculate all possible permutations (Hope, 1968). However, the precision of the P-value increases with the numbers of permutations. Generally, at least 1000 permutations should be done for tests with a α-level of 0.05 and at least 5000 permutations should be done for tests with an α-level of 0.01 (Manly, 1991; Anderson, 2001a). Statistical analysis and permutation tests are conducted with the software R using the package TraMineR for sequence analysis (Gabadinho et al., 2009).

3.5. Results

Univariate analysis of variance

The one-way analysis of variance confirmed that individuals with similar characteristics experience more homogeneous family trajectories. Table 3.1 shows the decomposition of variance for optimal matching distances. The results of the F-tests indicate a certain grade of homogeneity within groups. The P-values are inferior to 0.001 and suggest that the observed differences among groups are statistically significant. Among the background characteristics, one can notice that the two variables with bigger F-statistics are race/ethnicity and family composition. This suggests that these two variables are discriminant in explaining heterogeneity in life course. Among the geographical context, we can notice how median income and poverty rate exhibit great values in the F-statistics. This means that individuals who used to live (during Wave I) in neighborhoods with similar economic status are more likely to experience homogeneous family trajectories. These variables, in fact, contribute significantly to explain the total variability of the sample. These results support hypotheses 2 and 3.

Table 3.1.: One-way analysis of variance. Pseudo F-test, p-values based on 1,000 permutations

Variable	SS_B	*df*	SS_W	*df*	SS_{tot}	*df*	F-stat	P-value
Background characteristics								
Race/ethnicity	6785.55	3	428787.15	6851	435572.70	6854	36.14	< 0.001
Family composition	6982.25	3	429056.88	6858	436039.13	6861	37.20	< 0.001
Parents' education	7185.19	4	428853.94	6857	436039.13	6861	28.72	< 0.001
Income	5437.28	3	313328.67	5089	318765.95	5092	29.44	< 0.001
Religiosity	264.41	1	435774.72	6860	436039.13	6861	4.16	< 0.001
Geographical context								
School type	1741.62	1	434297.51	6860	436039.13	6861	27.51	< 0.001
Unemployment rate	2399.48	2	441704.42	6971	444103.91	6973	18.93	< 0.001
Median income	5728.38	2	438375.52	6971	444103.91	6973	45.55	< 0.001
Immigrant population	1698.08	2	442405.83	6971	444103.91	6973	13.38	< 0.001
Poverty rate	5052.13	2	439051.78	6971	444103.91	6973	40.11	< 0.001

Background characteristics

A simple decomposition of variance is sufficient to test if there are significant differences among groups. On the other hand, a simple test does not provide any indication of the differences among different categories. Table 3.2 and Figure 3.1 show the variance among the different categories of background variables. The total variance of the sample is 63.54 and represents a measure of the dissimilarity between a couple randomly chosen from the sample. Values above this level indicate that the variability is greater than the mean level of the sample, while inferior values gives indications of homogeneity. Race/ ethnicity shows great differences among categories. White and Asian women exhibit lower variability while Black and Hispanic women are more heterogeneous. Family composition indicates that women who used to live with both biological parents at wave I have more homogeneous family trajectories. On the other hand, women living in a single family or in other types of families have more differentiated family patterns. Parents' education and family income are associated with lower heterogeneity (Hypothesis 4). In particular, young women with college educated parents and those with parents in the last quartile of income distribution, exhibit lower levels of sequence variability. We do not observe, instead, substantial differences by religiosity.

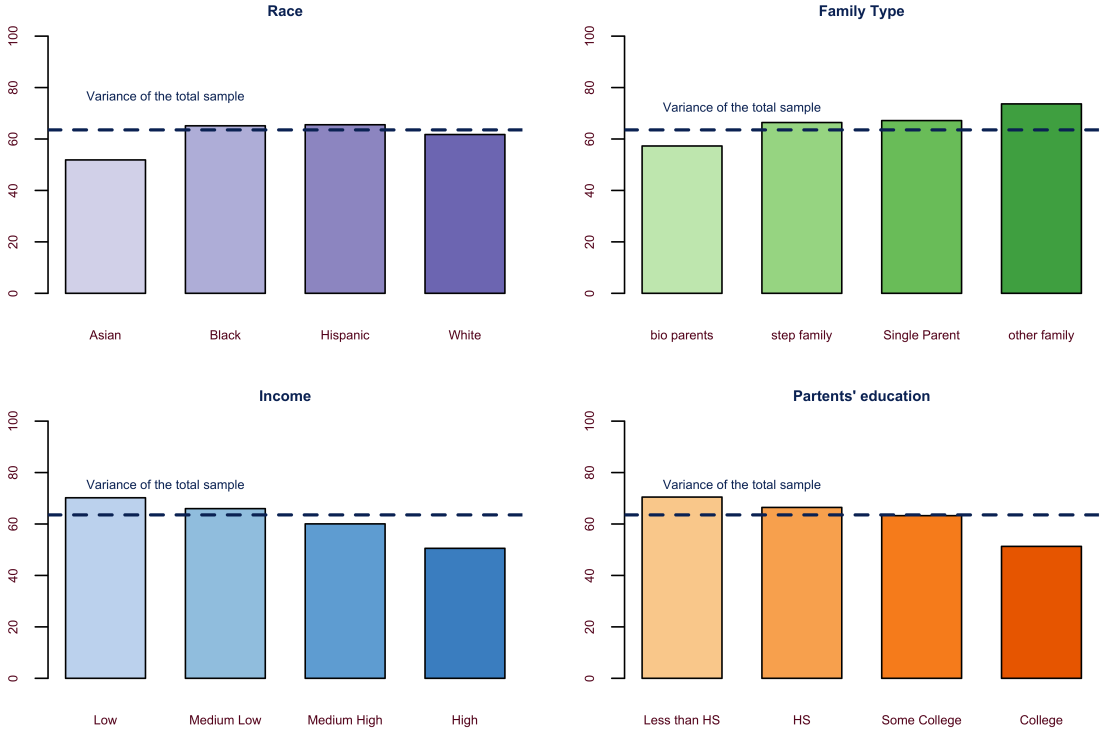
The analysis clearly shows that background characteristics contribute significantly to explain heterogeneity in life course. In particular, it seems that lower levels of heterogeneity are associated with higher social class, measured in terms of parents' education and family income. However, the contribution of each variable in explaining the total variability of life sequences is small. The R^2 is quite low for each variable tested in the model. Less of the 2% of the variance, in fact, can be explained by such characteristics as income or parents' education. The low levels of R^2 obtained with the decomposition of variance may indicate that there are many unobserved variables that contribute to explaining heterogeneity in life course. To test the hypothesis that background characteristics matter, I adopt a fixed effect approach. I assume, in fact, that most of the background characteristics are due to the family of origin and thus are common among

Table 3.2.: Analysis of Variance. Background characteristics

Variable	n	Variance	Pseudo R^2
Race/ethnicity			
Asian	365	51.87	0.016
Black	1488	65.13	
Hispanic	1083	65.54	
White	3919	61.74	
Family composition			
Living with biological parents	3523	57.28	0.016
Step family	696	66.41	
Single Parent	2102	67.16	
Other family	541	73.65	
Parents' education			
Less than high school	780	70.47	0.016
High school or equivalent	2112	66.44	
Some college	1203	63.24	
College or more	1762	51.30	
Unknown	1005	66.75	
Income			
Low	1117	70.20	0.017
Medium-low	1408	65.98	
Medium-high	1287	60.05	
High	1281	50.54	
Religiosity			
Never attend church	1891	63.40	< 0.001
Attend once a week religious services	4971	63.55	
Total sample	6862	63.54	

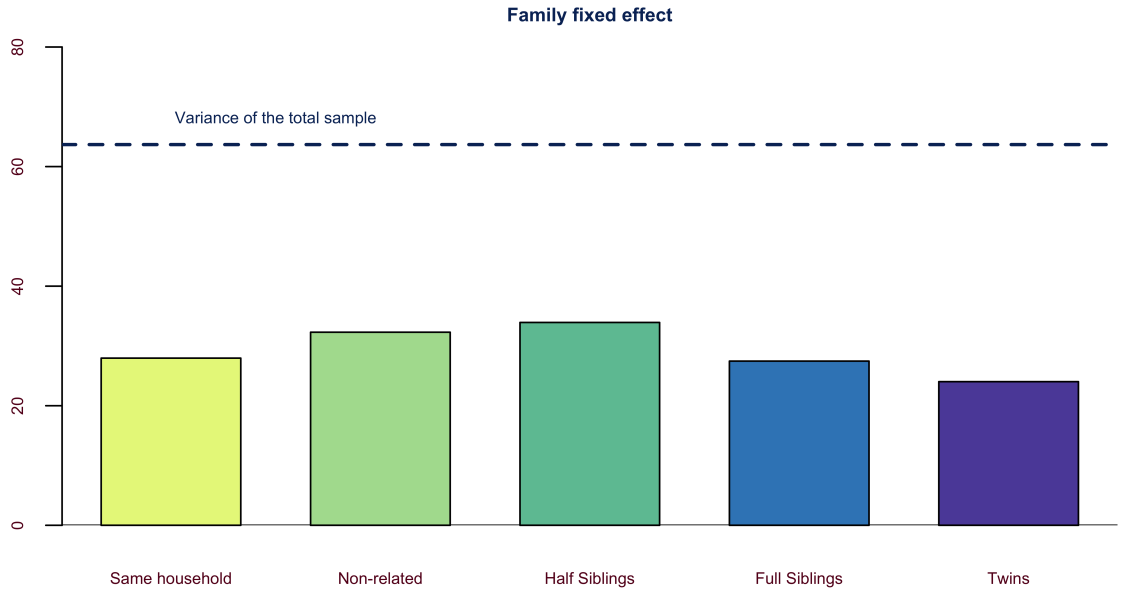
the members of the household. Therefore, I take the advantage of the pairs' sample of Add-health data, in order to test homogeneity within the same family. I hypothesize, in fact, that individuals in the same household experience more similar family trajectories, because of the influence of a common family environment. Moreover, I expect that the typology of family links matters. Full siblings may be more similar than half siblings or non related members (mostly cousins) because they have in common both biological parents. Twins may be more similar because they share same more of the same charac-

Figure 3.1.: Variability in family trajectories by background characteristics



teristics (including genetics). A descriptive analysis of similarity among family members is shown in Figure 3.2. The total number of individuals with at least another member of the household included in the sample was 1,956 (829 full siblings, 288 half siblings, 537 Twins, 396 Non-related members). Although the sample size of family pairs is low, the results show the expected dependency. Individuals in the same family have a high degree of similarity in family trajectories. Moreover, it is possible to notice that sequence similarity increases with the family relation. Twins are, in fact, more homogeneous than half siblings and full siblings. The mean distance between trajectories of individuals living in the same household is the 44% of the mean distance between couples randomly chosen in the sample. In particular, the average distance is slightly higher for half-siblings (53%) than full-siblings (43%) and twins (38%). These results confirm the background hypothesis (Hypotesis 1) since young women that have similar background characteristics are more likely to experience similar family trajectories.

Figure 3.2.: Family fixed effect. Variability in life trajectories among members of the same family. N=1,956



Geographical context

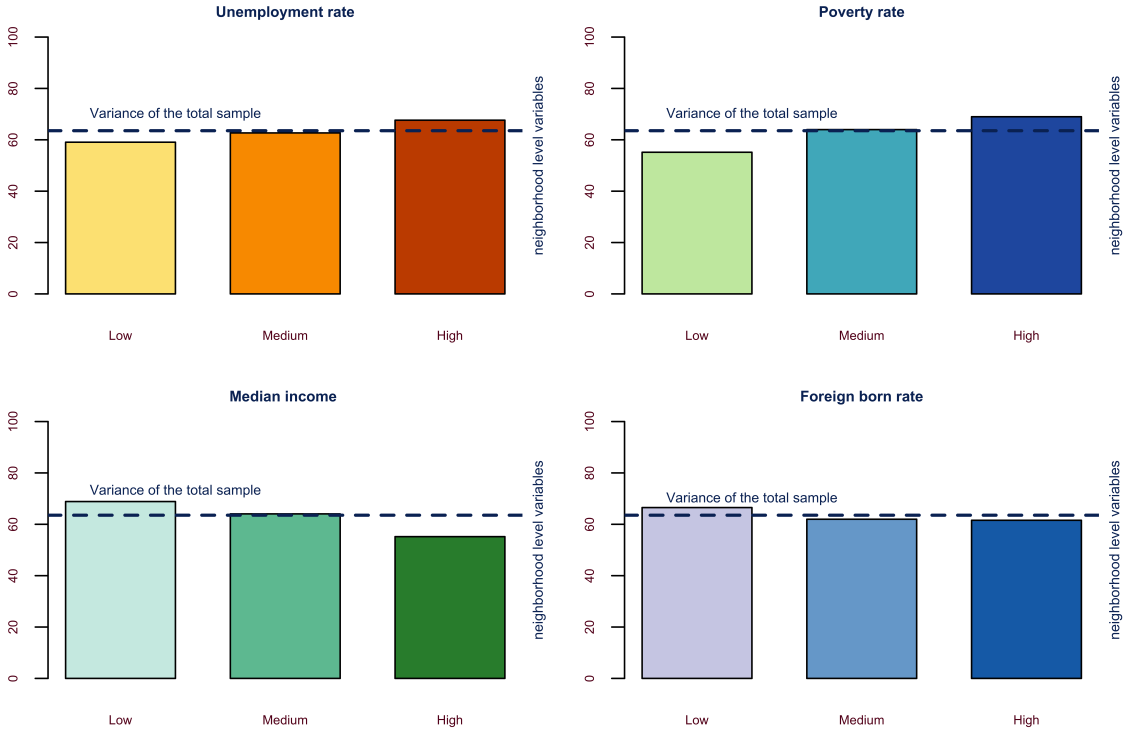
The geographical context in which individuals grow up may have an influence on future family trajectories. Characteristics such as the unemployment rate, poverty or the average income of a geographic area may give an indication of the social context of the neighborhood. The results shown in Table 3.3 and Figure 3.3 indicate that individuals living in neighborhoods characterized by low unemployment, higher income and low poverty rate experience more homogeneous family trajectories. Another important source of social capital is the school attended during adolescence. Peers can be influential in terms of behaviors (for example sexual initiation) and life course decisions. Furthermore, school represents an important place for the marriage market. It is, therefore, reasonable to expect that school significantly contributes to explaining life course variability. The results, indeed, show that the typology of school matters. Students attending private schools exhibit less variability in life course trajectories.

Table 3.3.: Variability in family trajectories by geographical characteristics. Block level variables

Variable	n	Variance	Pseudo R^2
School type			
Public	6400	64.37	0.004
Private	462	48.34	
Unemployment rate			
Low	2196	59.06	
Medium	2250	62.67	0.005
High	2528	67.64	
Median Income			
Low	2379	68.86	
Medium	2364	64.05	0.012
High	2231	55.19	
Foreign-born proportion			
Low	2444	66.52	
Medium	2352	61.97	0.004
High	2178	61.56	
Poverty rate			
Low	2215	55.15	
Medium	2274	63.97	0.011
High	2485	68.98	
Total sample	6862	63.54	

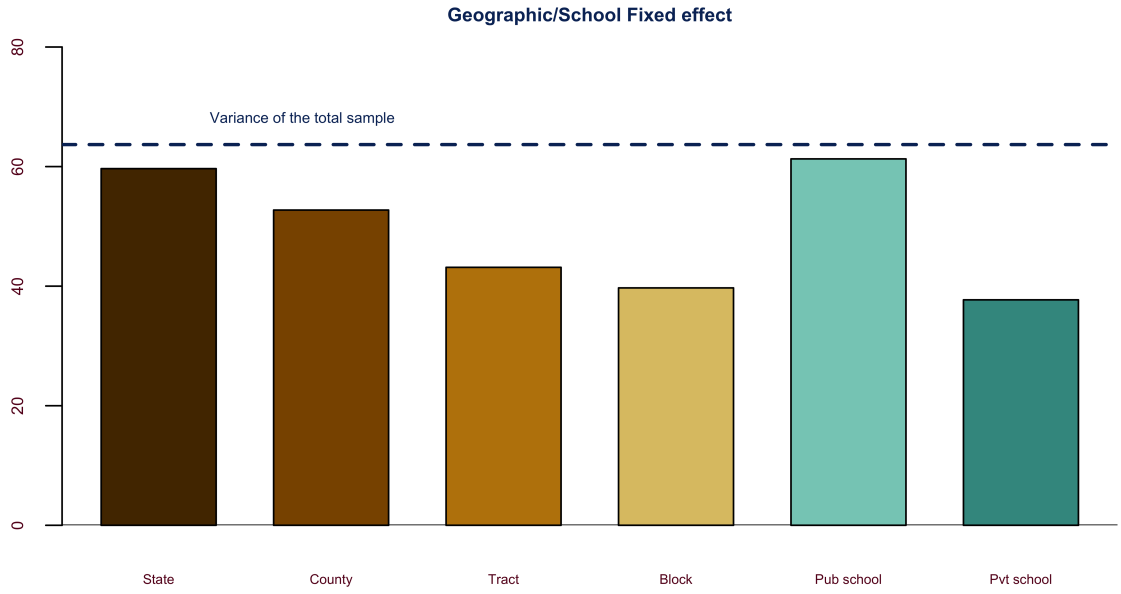
Analogously to the family characteristics, I proposed a fixed effect approach to test the influence of geographical area. Although the effect of variables measured at geographical level are statistically significant (see table 3.1 for statistical test of the pseudo F-statistic), the fraction of variance explained by contextual socio-economic variables is very low. This motivates a further analysis on the variability within individuals at different geographical levels. I suppose, in fact, that an increase in the geographical detail is associated with an higher level of homogeneity in life course. In Figure 3.4, I report the amount of variability among individuals in different contextual levels (34 states; 127 counties; 703 tracts and 1129 blocks). The Figure clearly shows that individuals who used to live in the same geographical area have more similar trajectories. The hypothesis is confirmed since the

Figure 3.3.: Variability in family trajectories by geographical characteristics. Block level variables



higher the geographical detail (state, county, tract and block) the more life course become similar. Results show that the average distance between individuals in the same state is the 94% of the average distance of the sample. Average distances decrease to 83% in the same county, 68% in the same tract and 62% in the same census block. The analysis by geographical levels is accompanied by a comparison of young women who attended the same high school (6,400 public schools and 462 private schools). It is interesting to notice the family trajectories of women who went to the same public school are slightly less heterogeneous than the average level. On the other hand, those who attended private schools have the highest level of homogeneity, more than women living in the same block (59% of the average distance). These results confirm the hypothesis 2 (*Geographical hypothesis 2*).

Figure 3.4.: Geographical fixed effect. Variability in life trajectories within the same geographical area. N=6,916



Multivariate models

Until now, I presented the results of univariate analyses. However, most of the variables taken into consideration are highly correlated and the interaction between them is not negligible. In addition, contextual variables are highly correlated with background since the place of living and the school attended by the respondents is supposedly the result of family choices. It is very likely, in fact, that low-educated families live in lower socio-economic neighborhoods and their children are more likely to attend public schools. For this reason, it is convenient to accompany univariate analyses with multivariate models where a set of variables are taken into consideration simultaneously. In table 3.4 the results of a multivariate analysis of variance are shown. The model has been selected through a backward procedure. Since race/ethnicity are highly correlated with most of the other variables, the model is presented separated by race/ethnicity. Results confirm that family composition and parents' education contribute to the explanation of family trajectories. The multivariate model confirms the results obtained in the univariate analyses. Family composition and parents' education contribute significantly to explain heterogeneity for all the racial/ethnic groups. Parents' birthplace is more relevant among

Asian women while religiosity appears to be relevant especially for White women. The only contextual variable that contributes to explaining heterogeneity in life course is poverty level. A possible explanation of the inclusion of only one contextual variable is that the variables included in the model are highly correlated. Overall, the amount of variance explained by the model is modest, since less than 5% of the total variability in family trajectories is explained by the model.

3.6. Discussion

In this chapter, I attempt to study the variability in family trajectories within a cohort of young US women. I propose to use sequence analysis and I present a non-parametric approach to test differences between groups of life course sequences. Although most of the research on variability in life course has been done by looking at historical trends or by doing cross-country comparison, the analysis of variation *within* a cohort can contribute to shed a light on social stratification in life course. The analysis is focused on two different sources of variability: background characteristics that include demographic variables and socio-economic status, and contextual variables. The analysis shows that both background and context matter. Among the background characteristics, race/ethnicity, family composition and parent's education are the variables that most contribute to explaining divergence in family trajectories. Among contextual variables, school typology and economic indicators of the neighborhood contribute significantly to explain distances of life course trajectories. These results, accompanied by a "fixed effect" analysis on variations within family and geographical areas, confirm hypotheses 1 and 2 *(Background and geographical context hypotheses).*

Once the first two hypotheses are confirmed, I attempt to investigate what are the categories that exhibit greater homogeneity in life course. Two complementary hypotheses concerning socioeconomic resources have been stated. Does socioeconomic status reduce or increase variability in life course? The results give evidence of the latter. Women with more educated parents and higher family income show less heterogeneity in life course. Also the ones who attended private schools, lived in high-class neighborhoods with less

Table 3.4.: Multivariate model of analysis of variance (MANOVA). Results based on 1,000 permutations.

	PseudoF	PseudoR2	p-value
	White n=3,919		
Living with biological parents at WI	32.39	0.007	***
Parents' education (college)	52.24	0.012	***
Parents born in US	4.02	0.001	***
Religiosity WI (once a week or more attend religious services)	23.39	0.005	***
Poverty level above median (neighborhood level)	27.83	0.006	***
Total	38.69	0.043	***
Black n=1,488			
Living with biological parents at WI	9.98	0.007	***
Parents' education (college)	7.91	0.005	***
Parents born in US	2.089	0.001	**
Religiosity WI (once a week or more attend religious services)	1.73	0.001	*
Poverty level above median (neighborhood level)	2.75	0.002	**
Total	6.071	0.0200	***
Hispanic n=1,083			
Living with biological parents at WI	9.075	0.008	***
Parents' education (college)	3.06	0.003	***
Parents born in US	4.48	0.004	***
Religiosity WI (once a week or more attend religious services)	1.82	0.002	*
Poverty level above median (neighborhood level)	2.64	0.002	**
Total	4.49	0.020	**
Asian n=365			
Living with biological parents at WI	2.36	0.006	**
Parents' education (college)	3.60	0.010	***
Parents born in US	7.07	0.02	***
Religiosity WI (once a week or more attend religious services)	1.33	0.004	
Poverty level above median (neighborhood level)	0.64	0.002	
Total	3.45	0.046	***

p-values: ***<0.01; **<0.05; *<0.1. Based on 1,000 Permutations

unemployment, higher median income and less proportion of people living in poverty exhibit greater similarity in family trajectories. The results confirm hypothesis 4 *(social*

class increases standardization) and indicate that social class is associated with standardization in family trajectories. Although we do not have indications on the mechanism that links standardization and socioeconomic status, it is reasonable to suppose that part of the standardization originates from a delay in family transitions due to a longer permanence in the education system. College attendance, in fact, is usually associated with later family transitions. On the other hand, family and educational trajectories are strongly correlated and one trajectory influences the other. College attendance delays the age at first union and age at childbearing and may be responsible of greater homogeneity in family trajectories. On the other hand, educational patterns are influenced by previous family preferences and expectations. Social class may increase standardization since women with greater socio-economic resources (in terms of of human capital and social capital) can achieved the desired family trajectories. On the other hand, more disadvantaged women are more likely to experience unexpected variations in family trajectories. This, at least during early adulthood, increases heterogeneity in life course.

The methodological contribution of this chapter is to present a technique that can be used to test differences between groups of life course sequences. The method presented is a generalization of ANOVA in the case of metric and semi-metric measures. As in standard ANOVA models, the objective of the analysis is to partition the observed variance into components based on different sources of variation. This methods was first used in ecology to test ecosystem dissimilarity and in the study of genetic relations. In this chapter, I propose to use the same methodology with Optimal Matching distances for life course analysis. This method presents some advantages and disadvantages. The first advantage is the ability to test differences between groups of life trajectories using the whole information of Optimal Matching (or other sequence analysis distances) without using to data reduction techniques. Usually, when we refer to sequence analysis we calculate Optimal Matching distances and we use multivariate techniques (typically cluster analysis) to derive homogeneous groups of life trajectories. Although this procedure is very powerful in detecting typical patterns of life course, it does not help to test if there exist differences between subgroups of population. Using the entire dissimilarity matrix, it is possible to use more information from the optimal matching calculations. On the

other hand, using this pseudo-anova procedure, we can only draw conclusions on the variations between trajectories but not on the actual patterns that particular groups are more likely to follow. For example, we can say that White women experience more similar trajectories than other racial groups but we cannot describe what their typical pattern is. The two methods are therefore complementary and can be adopted to describe different aspects of life course analysis. Another limitation is that we can only deal with categorical variables. As in standard ANOVA, we test differences among groups but we cannot model directly the effect of continuous variables. The empirical analysis presented in this chapter uses Optimal Matching distances as a starting point. However, this procedure can be extended without any problem to any kind of metric and semi-metric distance. For life course analysis, we can therefore use other distances that measure dissimilarity between life trajectories (i.e. Hamming distance, Longest Common Subsequences[5]). The procedure is rather simple and intuitive since it is an extension of a basic statistic technique. The permutation approach is totally non-parametric and based on the number of permutations used. The disadvantage, in the case of multivariate models, is that the number of variables increases the complexity of the model and may be computationally intensive. In the case of life course analysis, this method can be particularly effective to test differences between groups of sequences, but the analyses presented in this chapter show that only a small part of the total variance (less than 5%) between life course trajectories is actually explained by the models presented. Nevertheless, the analysis presented in this chapter constitutes one of the first attempts to explain the heterogeneity in family trajectories using sequence analyses. The results indicates that social stratification is highly correlated with life course trajectories and should be taken into account in the study of variability in life course.

[5]See section 2.4.1 for examples of other sequence analysis distances

4. Family trajectories and health. A life course perspective

4.1. Introduction

During the last decade, there has been an increasing interest in the relationship between marital status and health, (see e.g. Schoenborn, 2004; Waite and Bachrach, 2000; Wood et al., 2007; Koball et al., 2010). This is partially motivated by the recent changes in family behavior that have occurred in the United States and many other Western countries, i.e. increase in cohabitation, delay in marriage and rise of non marital childbearing (Cherlin, 2005; Schoen et al., 2007). Studies on the United States highlight the positive association between marriage and a various range of health outcomes for both men and women. Married adults are less likely to die in any given period than the unmarried (Lillard and Waite, 1993; Dupre et al., 2009), they also appear to have better mental health than their counterparts (Lamb et al., 2003; Horwitz and White, 1998; Soons and Kalmijn, 2009; Meadows, 2009) and they are less likely to engage in unhealthy behaviors (Duncan et al., 2006).

Most studies examine health differences by marital status in order to identify the causal effect of marriage. Generally, they compare health outcomes of married men and women versus unmarried (or cohabiting) people or they examine the effect of changes in marital status across life course (Nock, 1981). Only a limited number of studies adopts a complete life course perspective. The life course paradigm assumes that individuals, as human agents, build their future on the basis of the constraints and opportunities experienced in the past (Elder, 1994). The process is iterative and cumulative, since initial advantages or disadvantages often are amplified with time (Giele and Elder, 1998). Life courses are

embedded in different time and location and are affected by the social context in which individuals live. In addition, different life domains are strongly interdependent.

Elder (1985) observes that a trajectory can also be envisioned as a sequence of transitions that are enacted over time. A transition is a discrete life change or event within a trajectory (e.g., from single to married), whereas a trajectory is a sequence of linked states within a conceptually defined range of behavior or experience. Transitions are often accompanied by socially shared ceremonies and rituals, such as a graduation or a wedding ceremony, whereas a trajectory is a long-term pathway, with age-graded patterns of development in major social institutions such as education or family. In this way, the life course perspective emphasizes the ways in which transitions, pathways, and trajectories are socially organized. Moreover, transitions typically result in a change in status, social identity, and role involvement. Trajectories, however, are long-term patterns of stability and change and can include multiple transitions. Using longitudinal or retrospective data, family trajectories can be described by the complete sequence over time of union status, childbearing and eventually work status. Life course scholars stress the importance of the long term effects of trajectories (Soons et al., 2009), together with other characteristics of life history. Rather than investigating the contemporaneous association between marital status and wellbeing, life course analysis looks at the entire development of family history, i.e. the whole trajectory. Under this perspective, characteristics such as type, number and duration of unions, or the order of events may have an effect on later health outcomes (Peters and Liefbroer, 1997).

In this chapter, I investigate the role of family trajectory, i.e. the whole sequence of family events, during the life course of early adults in shaping their health outcomes. I jointly consider union formation and childbearing, since the two life domains are highly connected and their intersections may have an effect on health outcomes. This chapter is divided in two parts. First, I focus on transitions and investigate if changes in *timing* (when events happen), *quantum* (what and how many transitions) and *sequencing* (in what order) (Billari et al., 2006; Billari, 2005), have an effect on the health of young women. In the second part, I classify life course trajectories into six groups representing different ideal-types of family trajectories and I explore the association of these trajecto-

ries with health outcomes.

4.2. Theoretical and empirical background

According to the life course health development (LCHD) model, health is the result of a continuous process that develops over an individual's lifetime (Halfon and Hochstein, 2002). In the LCHD model, health is a consequence of multiple factors operating in nested genetic, biological, behavioral, social, and economic contexts. These contexts change as a person develops. Therefore, health is seen as an adaptive process, composed by multiple transactions between the contexts mentioned above (e.g., genetic, social) and the biobehavioral regulatory systems (e.g., neurological, endocrine) that define human functions (Halfon and Hochstein, 2002). In other words, health is not a static phenomenon. It develops over time and changes as a function of experience. The LCHD model suggests that a person's health takes on a trajectory that results from the cumulative influence of multiple risk and protective factors during life course. Health, in turn, is a multidimensional concept that encompasses a large array of measures, including behavioral, physical, and emotional outcomes.

The association between family transitions and health is well documented. Changes in the family structure may affect health in several ways. In particular, Wood et al. (2007) distinguish five different health dimensions: health behaviors, mental health, physical health and longevity, health care access and use, intergenerational health effects. In this chapter, I will only consider the first three dimensions. Using a sample of young women in the United States, I study the consequences of family trajectories on self-reported health, depression, drinking and smoking behaviors.

A large number of works demonstrates that married people are healthier, happier and less likely to engage in health threatening behaviors (for a review see Wood et al., 2007; Schoenborn, 2004). These potential benefits of marriage have influenced, at least in part, several US governmental initiatives in recent years that encourage and support marriage (Lichter et al., 2003; Acs, 2007). Consequently, this led to a debate on the effectiveness of pro-marriage policies among the scientific community, (McLanahan, 2007; Amato, 2007;

Nock, 2005).

In the literature, the benefits associated with marriage are generally called the "protection effects" of marriage (Waldron et al., 1996). In their review, Musick and Bumpass (2006) suggest four possible explanations: institutionalization, social roles, social support and commitment. Marriage is an institution where spouses have defined social roles both inside and outside the household (Gove, 1972; Ferree, 1990). Moreover, marriage is a source of social support. Spouses provide intimacy, companionship and daily interaction. At the same time, married people are connected to a larger network (e.g. friends, kin). This enlarges the social capital from which spouses can draw on in case of need. Last, the public nature of marriage strengthens commitment and facilitates joint long-term investments, including financial, role specialization and time spent in the care of young children. Commitment strengthens bonds between partners and serves as a barrier to exit. It is not clear, however, if these benefits are unique to marriage or whether they can be extended to other intimate relationship, particularly cohabitation. Evidences are mixed: Wu and Hart (2002) find no health effects of entering into marriage or cohabitation in Canada. Horwitz and White (1998) find differences in happiness, but no disadvantages in terms of depression. Musick and Bumpass (2006) examine several dimensions of wellbeing including psychological health, social ties and relationship quality and they do not find significative differences between married and cohabiters. In a comparative research using data from 30 european countries, Soons and Kalmijn (2009) find that the cohabitation gap (with respect to marriage) in wellbeing is associated with the degree of acceptance of non-marital unions in the society.

Although there is an extensive literature on the association between marital status and health outcomes, a number of issues motivates a life course perspective. First, the association between marriage and wellbeing may reflect preexisting conditions. Healthy individuals may be more likely to possess certain characteristics, such as higher earnings, emotional health, and physical attractiveness, that make them more desirable marriage partners than those in poor health. In contrast, those with poor mental or physical health may lack the energy and well-being necessary to find a spouse or a partner. Most

of the studies take into account selection issues using longitudinal data and controlling for the "individual effect". This is done generally using "fixed effect models" or "lagged dependent variable" regression, where the researcher can take in consideration selection controlling for previous outcomes. Although these statistical models take into account selection, they generally do not solve the problem of reverse causation. In some situations, health status may be the cause, rather than the effect of family transitions. For instance, once married, those who are less healthy may be less able to communicate and to participate in activities with their partner, or may have difficulties to contribute financially to the household, all of which may increase the likelihood of divorce.

Second, when data on marital status are collected in a longitudinal survey, we often ignore what happens between the time periods that are taken in consideration. Cohabitation and marriage are not mutually exclusive. In the United States, about half of young adults live with a partner before marrying. For some people, cohabitation is a prelude to marriage or a trial marriage. For others, a series of cohabiting relationships may be a long-term substitute for marriage (Cherlin, 2005). Although cohabitation has become common in the United States, it rarely lasts long. About half of cohabitation relationships end through marriage or a breakup within a year (Seltzer, 2004; Bumpass and Lu, 2000). If we consider only the change in marital status between the two waves of a longitudinal survey, we may ignore possible variations occurring in between. This may lead to considerable bias if the time between two data collection is sufficiently large. For instance, we may not distinguish between an individual married for the first time and another one who remarried after a separation. Also, since many married people experience cohabitation, it may be difficult to separate the causal effect of marriage. Does marriage have a different effect if it is preceded by cohabitation? In this case, does the time of exposure to premarital cohabitation matter?

Third, the majority of studies focus on union status without taking into consideration the link with other life domains. Union status is clearly connected with other events that happen during the life course. Having a child, leaving parental home, finishing school, starting to work are strictly connected with the probability to enter (or exit) a union. For example, a couple may decide to marry because of an unplanned pregnancy,

or they can decide to postpone marriage until she/he reaches economic independence. Since different domains are strictly interlaced, it may be difficult to identify the effect of a single event, such as marriage or entering a cohabitation. Other variables may confound the effect of family transitions. There may be, in fact, interactions between family events and background characteristics such as race, socio-economic status or social context. For instance, Harris et al. (2010) observed that early marriage by young adults does not have a protective effect for African Americans as observed for whites. Moreover, numerous studies show that individuals who marry at young age have higher risk of marital dissolution (Martin and Bumpass, 1989; Bumpass et al., 1991; Lehrer, 1988; Teachman, 2002).

Numerous studies try to investigate the causal link between divorce and premarital unions. Marital dissolution is higher among couple who experienced cohabitation. This negative effect is partially explained by self-selection (Lillard et al., 1995) and it is associated with the degree of acceptance of non-marital unions in the society (Liefbroer and Dourleijn, 2006). Moreover, Mazzuco (2009) found that the cohabitation length effect on duration of marriage is time varying, being close to zero for the first 2-3 years of cohabitation and rising considerably in the following years. Also low socioeconomic status may constitute a barrier to enter marriage (Edin and Reed, 2005; Schoen et al., 2009) and lead to other family transitions.

Last, standard analyses do not consider variations in timing, quantum and sequencing of life course trajectories. It is not clear, in fact, how changes in the structure of trajectories affect health outcomes later in life. Most researches, in fact, do not take into account when transitions occurs (timing), how many (quantum) and in what order they happen (sequencing). Transitions that occur in different periods of life may have a different effect on wellbeing. For instance, age at first union may be associated to health outcomes. Marriages at age 18 and 30 are qualitatively very different, indeed. At the same time, the sequence of events is relevant on the study of family life course. Does marriage have the same effect on health if it is preceded by the birth of a child? Evidence shows that unmarried mothers fare worse in the marriage market, because they have greater chances of partnering with poorly educated and unemployed men (Ermisch and

Pevalin, 2005). However, it is not clear if this increases the risk of having worse health outcomes. Last, trajectories may be very different in terms of complexity. Some individuals may experience a large number of transitions while others may not. Does stability in family trajectories affect health outcomes? Does the number of transitions matter? Some scholars argued that the overall structure of the life course has changed in profound ways, becoming "de-standardized," "de-institutionalized," and increasingly "individualized" (Macmillan, 2005; Shanahan, 2000; Elzinga and Liefbroer, 2007). It is not clear, however, what are the consequences of a de-standardization of family life course.

From a life course perspective, health outcomes are the result of the cumulative influence of multiple risks and protective factors experienced during the life course. For this reason the association between health and family formation should be expressed as an iterative process where health and family trajectories are mutually influenced. Under this perspective, it is necessary to take into account the whole trajectory in order to study the effects on health outcomes. The discussion above shows how difficult it may be to assess precise causal effects of family transition, unless the researcher relies on very strong assumptions. On the other hand, taking the whole trajectory as an input in statistical analysis is not straightforward (George, 2009). In this study, I use sequence analysis techniques to capture characteristics of the family trajectory such as complexity, sequencing and timing. Then, using Optimal Matching (Abbott and Tsay, 2000), I derived from data typical pathways of family formation using clustering techniques. Rather than identifying a causal effect of single family transitions, the aim of this paper is to explore associations between health outcomes and typologies of family trajectories. It may be possible, in fact, that certain typologies of family formation are associated with low health outcomes. This is relevant from a policy point of view. The study of family trajectories may highlight disadvantaged situations and it may permit to design appropriate interventions.

4.3. Contribution of the current study

The aim of this study is to explore the association between wellbeing and family trajectories from a life course perspective. In particular, I am interested in analyzing if there exists particular family trajectories associated with reduction in health status. To evaluate wellbeing I focus on the analysis of four different health outcomes: Self reported health, depression and risky behaviors (heavy drinking and smoking). I restrict the analysis to young women in age 30-33. I focused on young women for two reasons. First, the timing of family formation events tends to be earlier for women than for men. For example, the median age at first marriage in US is about 25 for women compared to 27 for men (Cherlin, 2004). Given the relatively young age of the sample I use, more women than men would have experienced family formation transitions. Second, becoming a parent is a central variable in this analysis, and men's reports of childbearing are less reliable than those of women. Indeed, one third to one half of men misreport non-marital births and births within previous marriages (Amato et al., 2008; Rendall et al., 1999).

A trajectory is defined as the monthly sequence of family states. The state-space is defined as follows. For every woman in the sample, I collect information about marriage and cohabiting relations. Moreover, I gather information about the age (in months) at first birth. The combination of union status with parenthood gives these six states: Single; Single Parent; Cohabiting; Cohabiting Parent; Married and Married Parent. Union states are reversible since from cohabitation it is possible to go into marriage or to return to single after a family disruption. Parenthood instead, is not reversible, i.e. from Single Parent a woman can only go to Cohabiting Parent or Married parent. The six states configuration follows the work by Schoen et al. (2007), where the authors examined early family transitions using a multi-state life table framework. The monthly detail permits to address in a precise way the order of transitions and to reduce the bias due to time interval. Differently from Amato et al. (2008), I take into consideration only family events (i.e. unions and childbearing) to focus on the relationship between health and family trajectories.

Following a life course perspective, I intend to analyze the association between differ-

ent types of family trajectories and self-reported health, depression symptoms and risky behaviors. In the first part of the empirical analysis, I focus separately on variations in timing, quantum and sequencing of family transitions. In the second part, I classify family trajectories in homogeneous groups sharing similar characteristics. The effect of selection and confounding variables is considered using appropriate statistical models. In reference to variation of timing quantum and sequences, I specify three different research hypotheses.

H1: Women who have earlier transitions have lower health outcomes. (Timing hypothesis)

I hypothesize that women who postpone family formation are more likely to invest in education and accumulate human capital. Young mothers or young women that enter an union have, in fact, less time to accumulate resources that contribute avoiding poor health and depression (Miech and Shanahan, 2000). Higher education also prevent women from engaging in behaviors that can damage their health. Furthermore, low educated women are more likely to match low educated men with higher probability of being unemployed and with lower income. Last, early marriage and early motherhood are associated with a higher probability of marital disruption that, in turn, is associated with major stress (Ermisch and Pevalin, 2005; O'Connell and Rogers, 1984).

H2: Women with "disordered" trajectories have lower health outcomes. (Quantum hypothesis)

Women who experience a large number of transitions are more likely to have less stable unions and may experience more traumas that can be dangerous for health development. The concept of "disorder" has been introduced for the first time by Rindfuss et al. (1987) in the study of transition to adulthood and parenthood. Individuals have expectations in terms of the role they assume in the society. A "disordered" life course may reflect difficulties to achieve the desired social role and fulfill the expectations. Also the lack of stability in family roles may be associated with more stress and less support from others. The "disorder" of life course is evaluated with a series of measures indicating the stability

of the trajectory.

H3: Women who have more non-normative transitions experience lower health outcomes. (Order hypothesis)
Family transitions are not qualitatively equivalent. I expect that family transitions that are recognized by the society as "normative" do not have negative effect on health. On the contrary, I expect that "non-normative" transitions are associated with lower outcomes. Individuals have expectations about the order of life-course events, even if sanctions are not applied. In fact, many sociological theories build in an expected sequencing of events in the transition to family. For example, first marriage is still sometimes equated with the beginning of exposure to the risk of parenthood. The variable ordering of events in the life course is a contingency of some importance in the life cycle (Hogan, 1978).

In the second part of the empirical analysis, I focus on family pathways. Since the possible combinations of family trajectories are enormous, I derive from data homogeneous clusters of trajectories. The resulting typologies of family pathways describe simultaneously different combination of timing, quantum, and sequencing. In analogy with Amato et al. (2008), I describe family formation using typical patterns of formation derived by empirical observations. The advantage of using classes is to reduce the (almost) unlimited number of combinations to a manageable number of groups that can be easily described. Differently form other studies (e.g. Amato et al., 2008), I am not interested in the precursors of different family pathways, but rather the consequences. Studying the health outcome of family typologies may help highlighting eventual disadvantages by subgroups of population.

4.4. Data and methods

4.4.1. Sample

The data I use come from Waves I and IV of the National Longitudinal Study of Adolescent Health (Add Health). Add Health is a longitudinal sample, nationally representative

of US adolescents who were in grades 7 through 12 in 1994-5. In the first wave, data were collected through in-home interviews with the adolescent participants and one of their parents. Typically, the parent interview was completed by the biological mother. Adolescents were interviewed again in a second wave one year later in 1996, again in a third wave collected in 2001-2002 and finally in a forth wave in 2008-2009. At the time of Wave IV, respondents ranged in age from 26 to 33 years. Since the goal of this study is to explore the implications of early life course trajectories, the sample is restricted to women who are 30 or older at Wave IV. Of this sample (n = 2,358), Wave IV weights are missing for 101 women. After dropping these cases, the final sample size is 2,259. At the time of the Wave IV data collection, 27% of women in the sample were 30 years of age, 54% were 31 years of age, and 19% were 32 years of age. Using retrospective questions from wave IV, I reconstructed the family biographies of women from age 15 to their age at wave IV.

Health outcomes

I created the following indicators to analyze different aspects of health status, with measures available both at Wave I and at Wave IV. Measures are expressed in a continuous scale, and indicate physical, mental health, drinking and smoking behaviors.

Self-reported Health

Status of current health was assessed with one question, "In general, how is your health?" (1= excellent, 2= very good, 3= good, 4= fair, 5=poor). Health status is therefore expressed in reverse order. Greater values indicate poor health status. I also report in the descriptive analysis the proportion of women reporting poor or fair health status (11% of the sample, Table 4.4).

Depression. A measure of depression has been constructed using questions from the CESD (Center for Epidemiologic Studies Depression) Scale (Radloff, 1977). In particular, nine questions out of this scale were asked (each based on the frequency of the event during the past seven days): bothered by things that usually dont bother you, couldnt shake off the blues, felt just as good as other people, had trouble keeping your mind on what you

were doing, felt depressed, felt too tired to do things, enjoyed life, felt sad, and felt that people disliked you (0 = never or rarely, 1 = sometimes, 2 = a lot of the time, and 3 = most of the time or all of the time). When appropriate, the coding was reversed so that high scores reflected high levels of depression. This indicator ranges from 0 to 21. I define as individuals with depression symptoms those who have a level of 9 or above (i.e. the ones who responded in average to have experience sometimes each of these symptoms, 18% - Table 4.4)

Smoking The number of cigarettes smoked in the last 30 days is used as a measure of smoking behavior. The percentage of women who report to have smoked at least an entire cigarette at wave IV is 27% (Table 4.4).

Heavy drinking. A scale of the frequency and severity of alcohol consumption has been created using this question: Within the last 12 months, on how many days did you drink five or more drinks in a row? Response options were 0 = never, 1 = one or two days, 2 = once per month or less, 3= two or three days per month, 4 = one or two days per week, 5= three to five days per week, and 6 = every day or almost every day. The resulting indicator is used as a continuous variable. Table 4.4 reports the proportion of respondents who had at least an episode of heavy drinking in the last 12 months (35% at wave IV).

Background characteristics

To control for compositional characteristics, I include in the models some indicators of demographic and socioeconomic status. Race/ethnicity is included: Hispanics, Black, Asian and White as a reference group. Parents' education is taken into account with a dummy variable indicating if at least one of the parent has college education. Also family composition at wave I is included. A dummy variable indicates if the respondent used to live with both biological parents during the first interview. Last, continuous values of age and age squared (measured in at Wave I) are included in the regression models.

4.4.2. Methods

In sequence analysis, life course trajectories are represented by monthly combination of union and childbearing states from age 15 to age 30. I define the state space to take

six possible values: Single (S); Single Parent (SP); Cohabiting (C); Cohabiting Parent (CP); Married (M) and Married Parent (MP). In sequence analysis, each life-course or trajectory is represented as a string of characters (also numerical), similar to the one used to code DNA molecules in the biological sciences. Thus, every trajectory is composed by a string of $(12) * 15 = 180$ values. The number of possible combinations is extremely large (6^{180}) and it is impossible to treat it with any statistical techniques. From a statistical point of view, sequences can be thought as the realization of a stochastic processes or alternatively as categorical time series. Life course sequences can be represented in several ways. A common approach is to describe the sequence with the state and its duration in time. For instance, an individual that stays single for 24 months, after that he has a cohabitation of 12 months and then she/he marries and stays married for 24 months can represented in this way:

(S, 24)-(C,12)-(M-24)

The sequence in the example describes the union status of a person for a period of five years.

Sequences differ in three dimensions: *timing*, *quantum*, and *ordering*. In this chapter, I attempt to define some basic indicators to measure variations in those three dimensions. The proposed indicators are then used in regression analysis to evaluate the association with health outcomes.

Timing

Timing refers to the duration of events, and specifically to the age at which different transitions happen in the life course. I propose three indicators for timing:

- Age at first transition (i.e., the earliest between first union and first child).
- Age at first union.
- Age at first child.

The three indicators are refereed to the period from age 15 to age 30. I only consider individuals who experienced the event by age 30. In Add-Health data, at age 30 the 94.4% of women exited singlehood, 93.6% experienced a union and 64.6% became mothers.

Quantum

Quantum indicates the number of events in a trajectory. I propose two indicators to evaluate the *quantum* of a sequence:

- Number of events from age 15 to 30.
- Sequences Turbulence.

The first is the number of transitions experienced from age 15 to 30 without distinguishing the type of transitions. The second is an indicator proposed by Elzinga and Liefbroer (2007) that measures the dynamics of a categorical time series. Turbulence takes into account, besides the number of transitions, the duration in different states. The turbulence index is, in fact, a composite measure of two aspects: variability in the time spent in different states and the number of distinct subsequences that can be extracted from the sequence. It gives an overall measure of the grade of disorder of a life trajectory (see e.g. Elzinga et al., 2008; Elzinga and Liefbroer, 2007; Widmer and Ritschard, 2009)

Sequencing

Sequencing indicates the order in which events happen in life sequence. I propose two indicators to evaluate the *order* in a family sequence:

- Number of normative transitions from age 15 to 30.
- Number of non-normative transitions from age 15 to 30.

I divide transitions in two groups: normative and non-normative transitions. Normative transitions are events in life course that are commonly accepted in the society (Rindfuss et al., 1987). In this study, I consider "normative" the sequence of events with this order: Single-Married-Married Parent. Each variation to this pattern is classified as "non-normative". It follows that: premarital childbearing, cohabitation, and any union disruptions are considered non-normative. The concept of normative is certainly

arbitrary and relative to the society in which the study takes place. Since long-term cohabitation in United States is not very common and marriage is still the primary form of union, I chose to include cohabitation on the list of non-normative transitions. Table 4.1 illustrates the classification of transitions.

Table 4.1.: Normative and non-normative transitions. Classification Table. 1="Normative"; 0="Non-normative"

	S_t	SP_t	M_t	MP_t	C_t	CP_t
S_{t-1}	[1]	0	1	0	0	0
SP_{t-1}	- [2]		-	1	-	0
M_{t-1}	0	0		1	0	0
MP_{t-1}	-	0	-		-	0
C_{t-1}	0	0	1	0		0
CP_{t-1}	-	0	-	1	1	

Regression Models

To examine the relation between the indicators above and health outcomes, I use regression models that take into account the effect of selection and confounding variables. The aim is to analyze if the change in the four outcomes between Wave I and Wave IV is imputable to some characteristics of family transitions. The time span between the two wave in consideration is around 15 years. In Wave I, the respondents are teenagers (age 13-16), while in the last wave they are 30-33 years old. This means that the two time periods considered represent two periods in life qualitatively very different. Health is a continuous process that develops across time. Health in early adulthood is very likely to be influenced by the level of health experienced in adolescence, childhood, infancy and during mother's pregnancy. Previous health levels, in turn, influence the family transitions. To account for this selection issues, I include in the model the previous level of health indicator as a regressor. To examine the impact of these indicators on health, I use a change (or lagged dependent variable) model that sets health at Wave IV as a function of the initial level of adolescent health at Wave I (Allison, 1990; Johnston, 1995). I then

[1] The empty diagonal indicates a permanence in the same state from time t to time $t+1$.
[2] The symbol $-$ indicates that the transition is not possible (i.e. from parenthood to singlehood).

include the characteristics of the trajectory, a set of time-invariant SES and control variables measured at Wave I. Such models can correctly be estimated as long as exogenous predictors are well controlled (Johnston, 2005).

The simple model is depicted in Equation (1)

$$Y_{i2} = \gamma D_i + \rho Y_{i1} + \beta X_{i1} + \epsilon_{i2} \tag{4.1}$$

Here, Y_{i2} represents a vector of health indicators measured at Wave IV (Time 2) for person i and Y_{i1} represents a vector of identical health measures at Wave I (Time 1). X_{i1} a vector of demographic controls and SES background at Wave I. The vector D_i represents the characteristics of the sequence from Wave I to Wave IV.

Alternatively we could assume that there is an individual effect such that $\epsilon_{i2} = \alpha_i + u_{i2}$ where α_i is the individual fixed effect and u_{it} a random shock. In this case we could use a fixed effect estimation, where the outcome is differentiated in order to drop the individual's time-invariant characteristics a_i (Angrist and Pischke, 2009). However, the fixed effect model is based on the presumption of time-invariant omitted variables. This assumption does not seem plausible since health is theorized as a development process that depends on many time-variant inputs that are not captured by the variables in the model. Also the time lag is sufficiently large (around 15 years). This avoids the risk that the time correlation explains all the variability in the outcomes.

Fixed effect (FE) appears to be particularly effective when we have information in small interval of time and we know changes in status. Using FE models, a change in status (i.e. marital status) can be associated with a change in the outcome. On the other hand, with a lagged dependent variable (LDV) strategy, we can include in the estimating equation time-invariant variables. While FE models control for time-invariant omitted variables, LDV model does not. In particular, this can lead to bias in the estimates if we attempt in identifying a causal effect of a treatment variable. However, in this case, the proposed estimation strategy seems to be a good compromise to give a portrait of the statistical association between trajectories and health outcomes.

Extracting typologies of life trajectories

The indicators proposed in the previous paragraph are useful to describe some characteristics of the life trajectory. However, they do not give any indication on the "type" of sequence. To describe completely family trajectories we need to study simultaneously *timing*, *quantum*, and *sequencing* in life course sequences (Billari, 2005). The complexity of life course suggests to adopt an holistic approach, where all the different components of the life course are taken into account. Abbott (1995) was the first to introduce sequence analysis in the social sciences using Optimal Matching algorithm (OM) as a method to compare different life sequences. This method has been used for the alignment of biosequences. The basic idea behind optimal matching is to measure the dissimilarity of two sequences by considering how much effort is required to transform one sequence into the other one. Transforming sequences entails three basic operations in this very elementary method:

- insertion
- deletion
- substitution

A specific cost can be assigned to each operation, and the total cost of applying a series of elementary operations can be computed as the sum of the costs of single operations. Thus, the distance between two sequences can be defined as the minimum cost of transforming one sequence into the other one. Hence, the resulting output is a symmetric matrix of pairwise distances that can be used for further statistical analysis, mainly multivariate analysis. Optimal Matching is a family of dissimilarity measures between sequences derived from the distance originally proposed in the field of information theory and computer science by Vladimir Levenshtein (Levenshtein, 1965), with the difference that in OM the three operations have different costs, (Lesnard, 2006). The choice of the operations' costs determines the matching procedure and influences the results obtained. This is a major concern about the use of this technique in social sciences (Wu, 2000). A common solution for assessing the substitution costs is to use the inverse of the transition

probability, in order to assign higher costs to the less common transitions (Piccarreta and Billari, 2007). I adopt this strategy in the empirical analysis.
Sequence analysis have been adopted in demography to study complex phenomenon in order to simultaneously study multiple demographic transitions (see e.g. Billari, 2001). Once obtained the dissimilarity matrix, we can apply standard reduction techniques to classify trajectories into homogeneous groups. The resulting groups are then used to describe "typical" patterns of transitions. Following the approach of McVicar and Anyadike-Danes (2002), I conduct a cluster analysis using Ward algorithm to identify six clusters of life sequences. Clusters can be described by choosing a representative sequence. Aassve et al. (2007) suggest to identify groups by using the medoid sequence, that is the sequence with the minimum distance from all of the other sequences in that cluster.

This group characterization of life sequences can be used as an input for further analysis, in particular regression analysis in order to explore the consequences of different life trajectories. For instance, Mouw (2005) uses the output of a clustering procedure as an input for a regression analysis under the heading "Does the sequence matter?" Regression analyses show important differences in the risk of experiencing outcomes such as poverty at age 35. Sequences are also found to influence subsequent happiness and depression status.

In this study, I analyze the consequences of family trajectories on health outcomes. I detect typical trajectories using cluster analysis on family sequences from age 15 to age 30. I only consider sequences from age 15-30 in order to have sequences of the same length for all the individuals. The resulting groups are then used as a categorical variable in a regression analysis. Using different "typologies" of trajectory allows to analyze the change in health status among different groups of individuals. This clustering procedure, for instance, allows to isolate the groups of single mothers who experience the birth of the first child outside a union, and do not experience stable union after childbearing. It is important, from a policy point of view to understand if any particular trajectory is associated with a decrease in health status. However, health status is measured at different ages for different individuals. This creates an asynchrony between the outcome

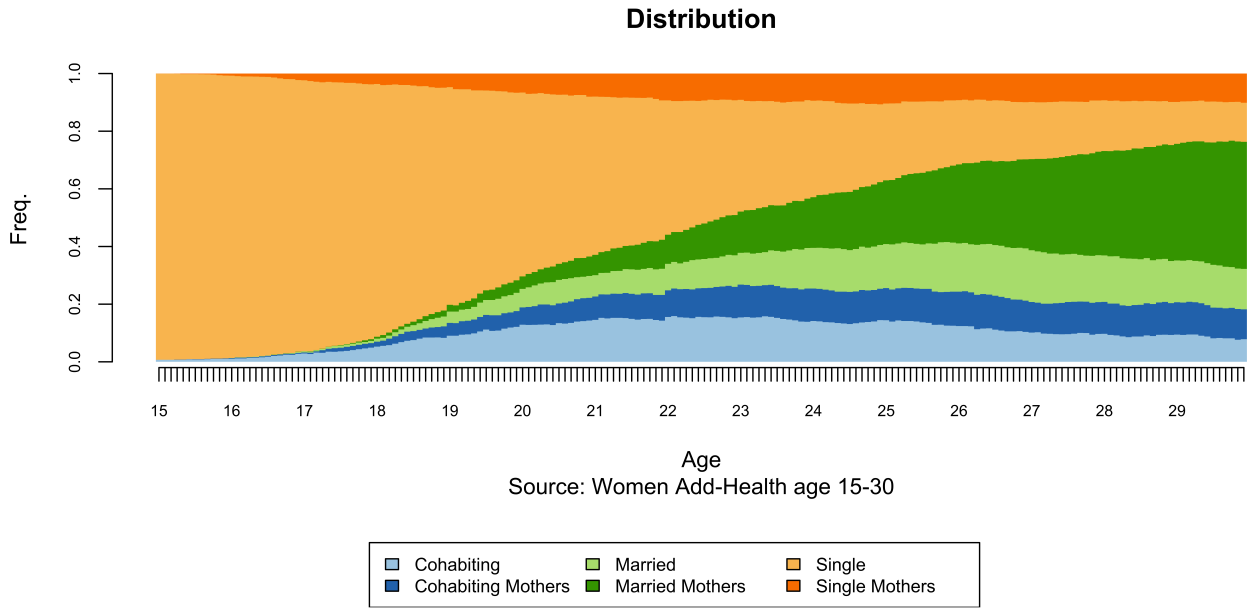
and the time used to describe the covariate. The ideal situation, would be to have individuals interviewed at the same age. To control for age effects I introduce age and age-squared in the estimation.

4.5. Analysis of trajectories

It is important to examine events in the initial years of early adulthood because the large-scale changes in cohabitation, marriage, and non marital fertility have particularly affected women in age 20-30. In terms of family transitions,those years are very "dense" (Rindfuss, 1991), with more demographic events occurring than during any other part of the life course. Figure 4.1 shows the distribution of "family states" from age 15 to 30. At age 30 very few women are single (because they did not enter an union, or because a disruption), the 55% are married and 18% are cohabiting (Table 4.2). Cohabitation is more frequent than marriage until age 23, then it slightly decreases at later ages. Motherhood increases with time, but it is predominant within marriage. The 44% of 30-years-old women are married and have at least a child (MP), while 11% are cohabiting mothers (CP) and the 11% are Single Mothers (SP). Only the 35% are childless and most of them are single.

The distribution of family states gives a picture of family states by age, but gives no indication about the dynamic of trajectories. Table 4.3 shows the most frequent trajectories observed among women 15-30. The representation in table 4.3 does not take into account the length of permanence in a state, but only the order of events. The first occurring pattern (11% of the sequences) includes cohabitation before marriage. The normative pattern of transitions is the second most common. Women that follow this pattern do not experience cohabitation. Only the fifth pattern contains individuals who do not experience any transition, while the sixth and the seventh indicates the presence of an union disruption. The first ten patterns cover 52% of all cases.

Figure 4.1.: Distribution of family states. Women age 15-30, weighted frequencies.



In table 4.4 I cross-classify health outcomes with some features of the sequences. Women who experienced marriage have better health outcomes. They are, in fact, less likely to report poor health, to suffer depression and adopt more healthy behaviors. On the contrary, women that have at least a cohabitation experience are more likely to have poor health. Furthermore, the proportion of smokers and heavy drinkers is greater among cohabiting and unmarried people. We do not observe great differences between mothers and non-mothers on self-reported health and depression. We observe, instead, differences in behaviors. In fact, mothers are less likely to be smokers or to drink than women who never had a child.

Although these descriptive tables show a relation between health and family status, the true impact can be masked by selection issues and by the effect of confounding variables. In table 4.5 I report the mean value of indicators of *timing*, *quantum*, and sequences for women conditional to their health status. Individuals with poor health status and depression symptoms have their first family transitions earlier than others. They usually experience more transitions, in particular the "non-normative" ones. Analogously, smoking and drinking behavior is associated with early exit from singlehood, younger age at

Table 4.2.: Weighted age percentage of women for marital status, cohabitation and motherhood from age 16 to age 30.

Age	Prop. married	Prop. cohabiting	Prop. with children
16	0.00	0.01	0.01
17	0.00	0.03	0.03
18	0.02	0.07	0.06
19	0.06	0.14	0.11
20	0.11	0.19	0.17
21	0.14	0.23	0.23
22	0.19	0.25	0.29
23	0.25	0.27	0.35
24	0.32	0.25	0.38
25	0.37	0.26	0.44
26	0.44	0.24	0.49
27	0.50	0.21	0.52
28	0.52	0.21	0.57
29	0.55	0.21	0.62
30	0.55	0.18	0.62

Table 4.3.: First 10 sequence pattern of transitions in Women 15-30. Weighted frequencies.

		Freq
1	S-C-M-MP	11.46
2	S-M-MP	10.46
3	S-C-M	5.93
4	S-C-CP-MP	4.41
5	S	4.37
6	S-C-S	3.46
7	S-C-S-C-M-MP	3.37
8	S-M	3.15
9	S-C	3.07
10	S-SP-CP-MP	2.77

Pattern representation indicates the sequence of events with durations ≥ 1

first union and first child and greater number of non-normative transitions.

Table 4.4.: Proportion of women in poor health, with depression symptoms, smoking and heavy drinking in the last 30 days. Frequencies by union status and motherhood.

	Prop. with poor health	Prop. with depression symptoms	Prop. smoking	Prop. drinking
Never Married	0.11	0.20	0.39	0.45
Ever Married	0.09	0.15	0.27	0.32
Never Cohabitation	0.07	0.17	0.16	0.20
Ever Cohabitation	0.10	0.17	0.36	0.42
Non-mothers	0.10	0.17	0.29	0.45
Mothers	0.10	0.17	0.32	0.32
Total	0.11	0.18	0.27	0.35

Table 4.5.: Indicators of *timing*, *quantum* and *sequencing* and health status.

	Poor Health		Depression		Smoking		Drinking		Total
	no	*yes*	*no*	*yes*	*no*	*yes*	*no*	*yes*	
Timing indicators									
Age at first transition	22.12	21.17	22.12	21.17	22.47	20.83	22.06	21.83	21.97
Age at first union	22.24	21.02	22.23	21.02	22.55	20.96	21.91	22.03	21.96
Age at first child	23.44	21.99	23.51	21.99	23.80	22.01	23.54	23.18	23.43
Quantum indicators									
Number of transition	3.09	3.40	3.11	3.41	2.9	3.69	3.08	3.36	3.18
Turbulence	6.43	6.48	6.45	6.48	6.27	6.89	6.42	6.68	6.52
Sequencing indicators									
Number of normative transition	1.08	0.97	1.10	0.97	1.12	0.93	1.19	0.96	1.10
Number of non-normative transition	2.02	2.43	2.01	2.44	1.80	2.80	1.897	2.40	2.08

4.5.1. Multivariate results

Early transitions have a negative effect on self-reported health and smoking behavior. Table 4.6 (and tables B.2,B.3 in the appendix) reports the results of the regression analysis. These results indicate that, controlling for previous health and compositional characteristics, transitions under age 18 are associated with poor self-reported health and increase in smoking. If we consider only union transitions or the age at first child, also transitions before age 20 are significantly different from transitions that happen later in life. Moreover, depression symptoms are associated with early childbearing. The dynamic of family trajectories has a similar effect. The number of transitions is associated with negative effect on self-reported health and smoking behavior. The more transition a woman experience between wave I and wave IV, the more she is likely to smoke and report poor health (see table 4.7). Other indicators of sequence dynamics, instead, do not show notable effect on health outcomes (see table B.4 in the appendix).

It is interesting to notice, however, what happens if we decompose the number of transitions into normative and non-normative (the distinction between normative transitions and non-normative is defined in table 4.1). Results in table 4.8 show that the two types of sequences have an opposite effect. While non-normative transitions have a negative effect on health outcomes, normative transitions are associated with less unhealthy behavior. Non-normative transitions are associated with a decrease in self-reported health and an increase in depression symptoms. Concerning smoking and drinking behaviors, we observe a protection effect given by normative transitions. Traditional family formation is therefore associated with reduction of risky behaviors. Controlling for other variables, non-normative transitions are associated with increase in the number of cigarette smoked and drinking occasions. Possible explanations are that non-normative transitions constitute major sources of stress. People who follow a normative path, instead, receive bigger support from friends and family.

The estimate results in tables 4.6,4.7,4.8 show similar levels of correlation between health outcomes in Wave I and Wave IV. The inclusion of lagged dependent variable

allows to take into account selection issues. I also included in the models' background variables indicating race composition, socio-economic status and the family composition at the beginning of the transition. Although previous health outcomes control for health selection, I assume that background characteristics can affect the level of health at Wave IV net of previous health outcomes. Estimates show that women with college educated parents have lower health outcomes and minor propensity to smoke. The propensity to engage in risky behavior changes with race. Black and Hispanic girls tend to smoke and drink less than their white counterpart. Moreover, African American women have a general tendency to report minor levels of health. Overall, these results show that women that move away from a traditional pattern have bigger risk to report poor health and above all to engage in risky behaviors. Therefore, these results show that *timing*, *quantum* and *sequencing* are important factors in the study of family formation.

4.6. Typologies of family trajectories

The analyses presented in the previous section show that women who move away from a "normative" model (especially in terms of age at first transition and order of events) are the ones who experienced greater decline on health status. Poor health outcomes are associated with early transitions, high numbers of changes in family status, and "non-normative" order of events. Traditional transitions seem to have instead a protective effect, especially on behavior.

Any how, previous analysis do not permit to identify what type of family patterns are associated with changes in health status. From a policy point of view, we are interested in detecting what subgroups of population risk more to experience poor health, for example, single motherhood (Furstenberg, 2005, 1998, 1976). Previous studies show lower levels of health among single mothers, in particular mental health (Cairney et al., 2003), propensity to smoke (Francesconi et al., 2010), and also higher level of mortality, (Mirowsky, 2005). Therefore, it is relevant to study the consequences of different patterns in family formation.

Table 4.6.: Regression estimates. Effects of timing indicators on health outcomes: age at first transition

	(1) Poor Health	(2) Depression	(3) Smoking	(4) Drinking
Age at first transition > 25 (ref.)				
Age at first transition < 18	0.270**	0.295	3.278***	0.0697
	(0.102)	(0.496)	(0.871)	(0.102)
Age at first transition 18-20	0.128	0.0691	1.771	0.0525
	(0.107)	(0.524)	(0.916)	(0.107)
Age at first transition 20-25	−0.0214	−0.0373	0.761	0.100
	(0.104)	(0.503)	(0.841)	(0.104)
Age at wave I	0.683	−4.167	−12.23	0.866
	(0.904)	(5.361)	(11.98)	(0.666)
Age squared at wave I	−0.0199	0.112	0.310	−0.0268
	(0.0251)	(0.147)	(0.329)	(0.0180)
Living with bio-parents at wave I	−0.0383	−0.185	−1.082	−0.00895
	(0.0533)	(0.250)	(0.640)	(0.0738)
College educated parents	−0.230***	−0.717**	−3.151***	0.0685
	(0.0636)	(0.273)	(0.771)	(0.0949)
Hispanic	0.184*	−0.112	−3.544***	−0.150
	(0.0905)	(0.437)	(0.755)	(0.0943)
Black	0.176**	0.567	−1.002	−0.278***
	(0.0618)	(0.342)	(0.770)	(0.0795)
Asian	0.208	0.0783	−1.643	−0.0998
	(0.120)	(0.389)	(1.242)	(0.155)
Self-reported health at wave I	0.270***			
	(0.0289)			
Depression WI		0.277***		
		(0.0259)		
Smoking WI			0.509***	
			(0.0318)	
Drinking WI				0.177***
				(0.0275)
Constant	−4.230	42.61	123.8	−6.337
	(8.132)	(48.88)	(109.3)	(6.181)
Observations	2255	2237	2237	2248
Adjusted R^2	0.141	0.108	0.310	0.060

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The number of possible combinations of sequences in family formation is almost unlimited. It follows that a convenient empirical strategy aims to reduce all the possible trajectories to a more manageable number. I used a cluster analysis to specify six groups of trajectories as representative of the entire set of sequences. The details of the analysis are presented in the Appendix. Below, I present a description of the sequences in each group, additional details can be found in table 4.9 and figures 4.2 and B.1. Clusters can also be described using their medoid sequences (Aassve et al., 2007). A medoid is the observation with the minimum distance from other individuals in a cluster. The advantage of using medoid sequences is to define the cluster using a real sequence that best represents the groups.

1. **Married mothers** *(S,73)(C,11)(M,12)(MP,84); n=693.* This is the largest group in the sample (29%). It is composed by women that follow a more traditional pattern, i.e. Single-Married-Married Mothers. Almost all of them experience both marriage and motherhood. Cohabitation is not rare, but generally short. Women in this class start family transition earlier than women in other groups (with the exception of single and cohabiting mothers). Although the number of transitions is comparable with the other groups, the number of "non-normative" transitions is limited.

2. **Late transitions** *(S,168)(C,12); n=648.* This group represents women that start family transition very late or the ones who have not experienced any transition by age 30. They stay single for the majority of the sequence and they eventually experience a transition to cohabitation. Very few of them are married or have a child by age 30.

3. **Married women without children** *(S,97)(C,12)(M,71); n=315.* This group differs from group 1 essentially for two reasons. Women in this group begin the family transition later and they remain longer married without a child. The average time in which they stay married without children (M) is 2 years and half, compared to 1 year in group 1. The result is that the majority of women in this group postpones childbearing after age 30. The majority of transitions is traditional

and cohabitation is generally short. Above all, this group is characterized by a postponement of traditional pattern.

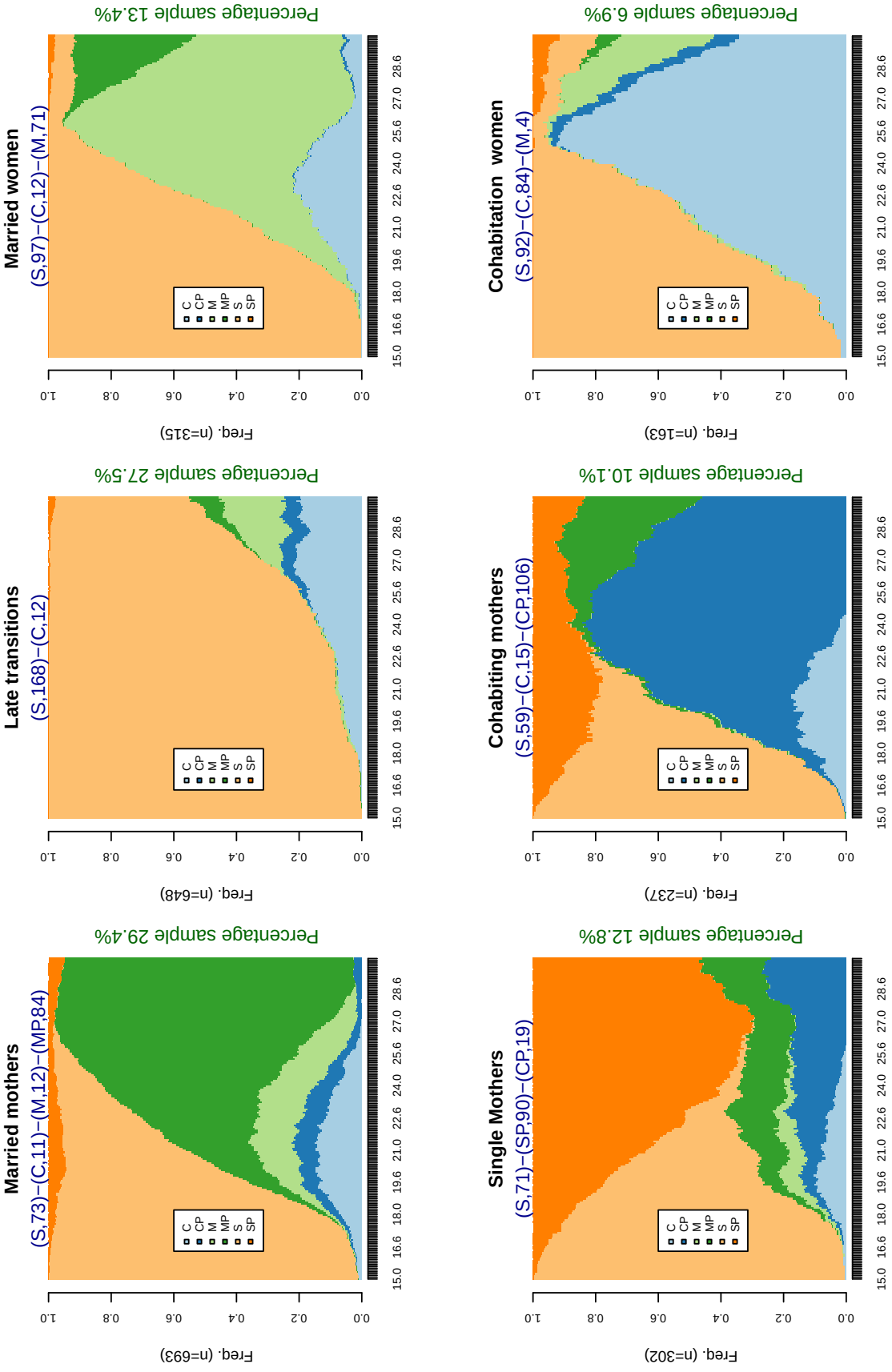
4. **Single Mothers** *(S,71)(SP,90)(CP,19); n=302.* This group identifies women who became mothers without being in an partnership. The group is characterized by very early transition to motherhood. Although there are some experiences of cohabitation, most of the time is spent outside a union. Women in this group experience in average more transitions than women in other groups. The majority of transitions are non-traditional. Single mothers are more likely to experience more than one cohabitation union.

5. **Cohabiting mothers** *(S,59)(C,15)(CP,106); n=237.* Women in this group differ from single mothers mainly for the fact that childbearing occurs during a cohabitation. This group is characterized by early transitions both to union and to motherhood. Similarly to single mothers, they experience a large number of transitions, most of them "non-normative" transitions.

6. **Cohabitating women** *(S,92)(C,84)(M,4); n=163.* The last group is characterized by cohabitation. It accounts for roughly 7% of women in the sample. Trajectories in this class are similar to group 2 (late transitions), with the difference that women in this group anticipate union to enter a cohabitation. The number of transitions is relatively low. Childbearing is postponed to later age.

Groups differ for compositional characteristics, in particular race composition and socioeconomic status (see table 4.9). Groups 4 and 5 have a higher proportion of African American women. These two groups seem to be the more disadvantaged in terms of family resources. Their families' income is noticeably inferior and a great proportion of them was not living with two biological parents at Wave I. On the contrary, women in the groups 2 and 3 seem to be more advantaged in terms of family income, education and family composition.

Figure 4.2.: Distribution of states



Single mothers, cohabiting mothers and cohabiting women (groups 3,4 and 6) report inferior level of health at Wave IV (table 4.6). The same groups also have higher probability to incur depression symptoms. This is partially explained by selection, since the same groups also have lower levels of health during wave I. Single and cohabiting mothers have a greater propensity to smoke at Wave IV. Drinking behavior, instead, is more frequent among cohabiting women and women who experience late transitions. Although we observe a general reduction in smoking from adolescent to adulthood, women who postpone family transitions (group 2) are the ones who have the biggest decrease.

To investigate the relation between health and family trajectories, I applied the same estimation strategy used in the previous section. Since family trajectories are subject to selection issues and confounding variables, I control for previous health outcomes (Wave I) and compositional characteristics in the regression models. The choice of the family pattern is very likely to be influenced by variables that are omitted in the regression model. Also the effect of reverse causation may not be negligible. On the other hand, the dependent variable is only a representation of a variety of trajectories and it cannot be thought as a treatment that is randomly assigned to the population. For this reason, the estimation results presented in table 4.11 only indicate a statistical association and they not have a causal significance. Nevertheless, results show some interesting aspects of the relation between health and family formation.

First, both women who have a child in early age and the ones who cohabit without children have lower self-reported health. On the other hand, women with a traditional pattern do not differ significantly to women who postpone family transitions. Second, cohabiting mothers are more likely to experience depression symptoms compared to other groups. Although single mothers are similar in many aspects, they do not differ from the reference group. A possible explanation is that depression is associated with the cohabiting experience, or in other terms with union instability. Last, smoking and drinking behaviors appear to be strongly influenced by family patterns. Trajectories with marriage seem to have a protective effect on the risky behaviors of women. Controlling for other variables, women of group 1 and 2 have lower probability to engage in heavy drinking

behavior, while women of group 2 experience a sensible reduction on the average number of cigarette smoked. This is consistent with other studies that show how marriage has a strong incentive on reducing risky behaviors (Duncan et al., 2006). However, it would be interesting to understand if this protective effect remains constant in time or if it has only a temporary effect. Overall, parents education has a positive effect on health - both physical and mental - and a reduction on cigarette smoking. Race has a mixed effect. Black women report less perceived health levels, but at the same time are less likely to engage in drinking behavior.

4.7. Discussion

Health is the result of a continuous process that develops over an individual's lifetime. Health trajectories are the consequence of a multitude of factors coming from genetic, biological, behavioral, social and economic contexts. Previous studies indicate that health is certainly connected with family events occurring during life course. Following the approach of Giele and Elder (1998), I distinguish between transitions (changes in family status) and trajectories (the whole sequence of transitions) in order to study jointly union formations and childbearing. Although the study of the dynamic inter-relationship between health and the life course has recently been an emerging topic, there is no general agreement on how trajectories should be conceptualized and analyzed. In this chapter, I use sequence analysis to describe life course trajectories. Describing family biographies as sequences of family states allows to analyze different dimensions of life course. In particular, I am interested in examining if there is a direct effect of *timing*, *quantum* and *sequencing* on health outcomes for young women. It emerges that, controlling for selection and background characteristics, changes in these dimensions affect health status. Early transitions have negative repercussions on self-reported health and smoking behavior (*hypothesis 1*). Although the experience of a large number of transitions is associated with negative effects (*hypothesis 2*), some particular transitions have a protective effect. Normative transitions (i.e. traditional unions, childbearing after marriage) have protective effects on behaviors (*hypothesis 3*). Women with numerous normative transi-

tions, in fact, smoke less cigarettes and have less occasions of heavy drinking. Above all, the indicators proposed indicate that sequence characteristics matters. In particular, it seems that moving away from normative family patterns (in terms of age-roles and order of events) is associated with a decrease in wellbeing.

In the second part of the chapter, I examine the consequences of different typology of trajectories. I individuate six classes representing typical patterns of family formation. Differences in terms of wellbeing and propensity to risky behaviors are substantial. Once controlled for selection and background characteristics, these differences are attenuated but still significant. Empirical results show that women with short experiences of cohabitation and women with a traditional pattern do not differ significantly to women who postpone family transitions. On the other hand, early childbearing and long cohabitation are associated with poor health status. Moreover, married women are less likely to smoke and to drink. These analyses partially confirm previous studies, in particular regarding the "protection effect" of marriage. Although selection and social background play most of the role, we still observe negative outcomes for women who experience early childbearing. We do not find much differences, instead, between single mothers and young mothers that have a child during a cohabitation.

Results show that early childbearing is associated with worse health outcomes. It is possible that women who anticipate motherhood have less resources (in terms of human and social capital) to tackle the stress of raising a child (especially if without a stable partner). Another complementary explanation is that early mothers are disadvantaged in the marriage market and they have difficulties to match with good men. Our results also show that married women are less likely to smoke and drink, confirming a "protection effect" of marriage. Cohabitation seems to have no negative effect if short and followed by a marriage. On the other hand, it is associated to poor outcomes (especially propensity to smoking and drinking) when it is persistent and accompanied by motherhood. It is possible, in fact, that short cohabitation, when followed by marriage, are becoming more and more accepted in the society. The aim of this paper is mainly descriptive. The mechanism of these relations, in fact, is beyond the scope of this work. Nevertheless,

these results, give evidences that family trajectories matter.

It would be interesting in the future, to investigate if these differences persist during the life course to see if the more disadvantaged groups are able to catch up with the others. Another open issue is the interaction between family transitions and social class. It may be, in fact, that family trajectories have different effects according to the socio-economic status of the family of origin. For example, the risk associated with non-normative transitions may not affect women coming from higher social class. Last, this study only deals with young women and ignores men. Comparing the trajectories of partners might help to understand the effect of previous family transitions in the marriage market. Any how, this study represents one of the first tentative to study the association between health and family formation using a life course perspective.

Table 4.7.: Regression estimates. Effects of quantum indicators on health outcomes: number of transitions

	(1) Poor Health	(2) Depression	(3) Smoking	(4) Drinking
Number of transitions	0.0457**	0.122	0.674***	0.0151
	(0.0149)	(0.0748)	(0.183)	(0.0192)
Age at wave I	0.325	−12.63*	−30.34*	0.529
	(1.475)	(5.981)	(14.84)	(1.113)
Age squared at wave I	−0.0102	0.343*	0.805*	−0.0175
	(0.0406)	(0.164)	(0.405)	(0.0299)
Living with bio-parents at wave I	−0.0622	−0.184	−1.215	−0.00217
	(0.0537)	(0.253)	(0.654)	(0.0736)
College educated parents	−0.266***	−0.739**	−3.410***	0.0735
	(0.0642)	(0.267)	(0.758)	(0.0920)
Hispanic	0.182*	−0.0408	−3.426***	−0.144
	(0.0917)	(0.435)	(0.747)	(0.0953)
Black	0.175**	0.579	−0.963	−0.274***
	(0.0619)	(0.341)	(0.783)	(0.0813)
Asian	0.198	0.0574	−1.785	−0.102
	(0.117)	(0.377)	(1.252)	(0.156)
Self-reported health at wave I	0.277***			
	(0.0292)			
Depression WI		0.273***		
		(0.0255)		
Smoking WI			0.508***	
			(0.0321)	
Drinking WI				0.176***
				(0.0273)
Constant	−0.915	119.5*	289.0*	−3.257
	(13.40)	(54.53)	(135.9)	(10.36)
Observations	2254	2236	2236	2247
Adjusted R^2	0.133	0.113	0.312	0.061

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.8.: Regression estimates. Effects of sequencing indicators on health outcomes: number of normative and non-normative transitions

	(1) Poor Health	(2) Depression	(3) Smoking	(4) Drinking
Number of normative transitions	−0.0180	−0.133	−1.122**	−0.216***
	(0.0315)	(0.144)	(0.354)	(0.0430)
Number non-normative transitions	0.0556***	0.160*	0.961***	0.0513*
	(0.0153)	(0.0769)	(0.205)	(0.0203)
Age at wave I	0.432	−12.21*	−26.79*	0.988
	(1.460)	(5.933)	(13.40)	(1.105)
Age squared at wave I	−0.0131	0.332*	0.710	−0.0298
	(0.0401)	(0.162)	(0.364)	(0.0297)
Living with bio-parents at wave I	−0.0459	−0.115	−0.796	0.0575
	(0.0540)	(0.253)	(0.657)	(0.0740)
College educated parents	−0.266***	−0.739**	−3.440***	0.0731
	(0.0638)	(0.266)	(0.745)	(0.0883)
Hispanic	0.163	−0.121	−4.121***	−0.216*
	(0.0912)	(0.435)	(0.752)	(0.0994)
Black	0.136*	0.425	−2.251**	−0.425***
	(0.0640)	(0.346)	(0.815)	(0.0862)
Asian	0.195	0.0529	−1.984	−0.124
	(0.117)	(0.384)	(1.207)	(0.144)
Self-reported health at wave I	0.273***			
	(0.0290)			
Depression WI		0.272***		
		(0.0256)		
Smoking WI			0.486***	
			(0.0327)	
Drinking WI				0.158***
				(0.0271)
Constant	−1.861	115.8*	257.3*	−7.337
	(13.27)	(54.10)	(123.1)	(10.26)
Observations	2254	2236	2236	2247
Adjusted R^2	0.136	0.116	0.329	0.090

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.9.: Descriptive statistics of typical group of sequences

	Married mothers	Late transitions	Married women	Single Mothers	Cohabiting mothers	Cohabiting women
			Union status and parenthood			
Ever married	1.00	0.37	1.00	0.51	0.44	0.44
Ever cohabited	0.70	0.71	0.72	0.83	1.00	1.00
Children	1.00	0.19	0.40	1.00	1.00	0.21
			Age at first transitions			
Age at first transition <18	0.58	0.12	0.32	0.63	0.79	0.46
Age at first transition 19-22	0.21	0.09	0.26	0.26	0.19	0.19
Age at first transition 23-25	0.21	0.38	0.42	0.11	0.02	0.36
Age at first transition >25	0.00	0.41	0.00	0.00	0.00	0.00
			Quantum and sequencing indicators			
Number of transitions Weave I-IV	3.37	2.41	3.14	3.89	3.79	3.32
Normative transitions	1.78	0.53	1.60	0.74	0.54	0.65
Non-normative transitions	1.59	1.88	1.54	3.15	3.25	2.67
			Compositional characteristics			
Proportion Black	0.1	0.18	0.06	0.34	0.31	0.14
Parents with college degree	0.19	0.27	0.38	0.15	0.07	0.22
Living with parents	0.49	0.56	0.63	0.29	0.26	0.52
Income family W1 (thousands of dollars)	41.54	51.92	53.52	33.38	34.59	41.73
Sex before 16	0.38	0.23	0.22	0.43	0.56	0.31

Table 4.10.: Descriptive statistics of typical group of sequences. Health outcomes.

	Married mothers	Late transitions	Married women	Single Mothers	Cohabiting mothers	Cohabiting women
			Health status at Weave I			
Prop. in poor health at WI	0.10	0.08	0.07	0.16	0.10	0.14
Prop. with depression symptoms at WI	0.25	0.23	0.22	0.28	0.29	0.35
Smoking at WI	0.39	0.42	0.35	0.42	0.50	0.46
Heavy drinking at Weave I	0.34	0.39	0.34	0.31	0.32	0.47
			Health status at Weave IV			
Prop. in poor health at WIV	0.09	0.08	0.10	0.13	0.12	0.14
Prop. with depression symptoms at WIV	0.16	0.15	0.13	0.17	0.26	0.23
Smoking at WIV	0.30	0.29	0.20	0.39	0.43	0.37
Heavy drinking at WIV	0.29	0.43	0.38	0.33	0.35	0.52

Table 4.11.: Regression estimates. Effects of family trajectories on health outcomes

	(1) Poor Health	(2) Depression	(3) Smoking	(4) Drinking
Late transitions (ref. category)				
Married mother	0.105	0.0705	−0.151	−0.311***
	(0.0694)	(0.301)	(0.768)	(0.0880)
Married women	0.0824	0.168	−1.906*	−0.225*
	(0.0784)	(0.389)	(0.913)	(0.109)
Single mothers	0.245**	−0.196	1.858	−0.162
	(0.0883)	(0.442)	(1.095)	(0.113)
Cohabiting mothers	0.211*	0.995*	2.243	−0.154
	(0.0919)	(0.498)	(1.153)	(0.123)
Cohabitation women	0.252*	0.887	0.327	0.330
	(0.118)	(0.481)	(1.824)	(0.230)
Age at wave I	0.650	−4.052	−12.08	1.173
	(0.876)	(5.366)	(11.90)	(0.645)
Age squared at wave I	−0.0193	0.108	0.301	−0.0354*
	(0.0244)	(0.147)	(0.326)	(0.0174)
Living with bio-parents at wave I	−0.0624	−0.210	−1.206	−0.0269
	(0.0536)	(0.248)	(0.645)	(0.0720)
College educated parents	−0.258***	−0.722**	−3.261***	0.0533
	(0.0643)	(0.274)	(0.772)	(0.0898)
Hispanic	0.152	−0.114	−4.162***	−0.182
	(0.0916)	(0.448)	(0.775)	(0.0980)
Black	0.126*	0.530	−1.945*	−0.324***
	(0.0630)	(0.350)	(0.815)	(0.0814)
Asian	0.198	−0.00107	−1.813	−0.134
	(0.121)	(0.394)	(1.185)	(0.149)
Self-reported health at wave I	0.277***			
	(0.0294)			
Depression WI		0.278***		
		(0.0261)		
Smoking WI			0.509***	
			(0.0318)	
Drinking WI				0.166***
				(0.0270)
Constant	−3.769	41.85	126.0	−8.809
	(7.889)	(48.93)	(108.6)	(5.973)
Observations	2255	2237	2237	2248
Adjusted R^2	0.132	0.113	0.308	0.077

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Bibliography

Aassve, A., F. C. Billari, and R. Piccarreta, 2007. Strings of Adulthood: A Sequence Analysis of Young British Women's Work-Family Trajectories. *European Journal of Population*, 23(3-4):369–388.

Abbott, A., 1995. Sequence analysis: new methods for old ideas. *Annual Review of Sociology*, 21(1):93–113.

Abbott, A. and A. Tsay, 2000. Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods & Research*, 29(1):3.

Acs, G., 2007. Can We Promote Child Well-Being by Promoting Marriage? *Journal of Marriage and Family*, 69(5):1326–1344.

Aisenbrey, S. and A. Fasang, 2010. New Life for Old Ideas: The "Second Wave" of Sequence Analysis Bringing the"Course"Back Into the Life Course. *Sociological Methods & Research*, 38(3):420–462.

Allison, P., 1990. Change Scores as Dependent Variables in Regression Analysis. *Sociological Methodology*, 20:93–114.

Amato, P., 2007. Strengthening marriage is an appropriate social policy goal. *Journal of Policy Analysis and Management*, 26(4):952–955.

Amato, P., N. Landale, and T. Havasevich-Brooks, 2008. Precursors of Young Women's Family Formation Pathways. *Journal of Marriage and Family*, 70:1271–1286.

Anderson, A., 2001a. Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian Journal of Fisheries and Aquatic Sciences*, 58(3):626–639.

Anderson, M., 2001b. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46.

Angrist, J. and J. Pischke, 2009. *Mostly harmless econometrics: an empiricist's companion*. Princeton NJ:Princeton University Press.

Bargeman, B., C. Joh, and H. Timmermans, 2002. Vacation behavior using a sequence alignment method. *Annals of Tourism Research*, 29(2):320–337.

Beath, K. J. and G. Z. Heller, 2009. Latent trajectory modelling of multivariate binary data. *Stat Model*, 9(3):199–213. doi:10.1177/1471082X0800900302.

Berge, J. M., M. Wall, K. W. Bauer, and D. Neumark-Sztainer, 2010. Parenting characteristics in the home environment and adolescent overweight: a latent class analysis. *Obesity (Silver Spring)*, 18(4):818–25. doi:10.1038/oby.2009.324.

Billari, F. C., 2001. The analysis of early life courses: complex descriptions of the transition to adulthood. *Journal of Population Research*, 18(2):119–142.

Billari, F. C., 2005. Life course analysis: two (complementary) cultures? Some reflections with examples from the analysis of the transition to adulthood. *Advances in Life Course Research*, 10:261–281.

Billari, F. C., J. Fürnkranz, and A. Prskawetz, 2006. Timing, Sequencing, and Quantum of Life Course Events: A Machine Learning Approach. *European Journal of Population*, 22(1):37–65.

Billari, F. C. and R. Piccarreta, 2005. Analyzing Demographic Life Courses through Sequence Analysis. *Mathematical Population Studies*, 12:81–106.

Blair-Loy, M., 1999. Career Patterns of Executive Women in Finance: An Optimal Matching Analysis. *The American Journal of Sociology*, 104(5):1346–1397.

Bruckers, L., J. Serroyen, G. Molenberghs, H. Slaets, and W. Goeyvaerts, 2010. Latent class analysis of persistent disturbing behaviour patients by using longitudinal profiles. *Journal of the Royal Statistical Society. Series C*, 59:495–512.

Brückner, H. and K. Mayer, 2005. De-Standardization of the Life Course: What it Might Mean? And if it Means Anything, Whether it Actually Took Place? *Advances in Life Course Research*, 9:27–54.

Brzinsky-Fay, C. and U. Kohler, 2010. New Developments in Sequence Analysis. *Sociological Methods Research*, 38(3):359–364.

Bumpass, L. and H.-H. Lu, 2000. Trends in Cohabitation and Implications for Children's Family Contexts in the United States. *Population Studies*, 54(1):29–41.

Bumpass, L., T. Martin, and J. Sweet, 1991. The impact of family background and early marital factors on marital disruption. *Journal of Family Issues*, 12(1):22–42.

Cairney, J., M. Boyle, D. Offord, and Y. Racine, 2003. Stress, social support and depression in single and married mothers. *Social Psychiatry and Psychiatric Epidemiology*, 38:442 – 449.

Cherlin, A., 2004. The Deinstitutionalization of American Marriage. *Journal of Marriage and Family*, 66(4):848–861.

Cherlin, A., 2005. American Marriage in the Early Twenty-First Century. *The Future of Children*, 15(2):33–55.

Clarke, K., 1993. Nonparametric multivariate analyses of changes in community structure. *Austral Journal of Ecology*, 18(1):117–143.

Clogg, C. C., 1995. Latent class models. In G. Arminger, C. C. Clogg, and M. E. Sobel, editors, *Handbook of statistical modeling for the social and behavioral sciences*. New York: Plenum.

Collins, L. and S. Wugalter, 1992. Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research*, 27(1):131–157.

Croudace, T., M. Jarvelin, M. Wadsworth, and P. Jones, 2003. Developmental typology of trajectories to nighttime bladder control: Epidemiologic application of longitudi-

nal latent class analysis. *American Journal of Epidemiology*, 157(9):834–842. doi: 10.1093/aje/kwg049.

Duncan, G., W. Bessie, and E. Paula, 2006. Cleaning Up Their Act: The Effects of Marriage and Cohabitation on Licit and Illicit Drug Use. *Demography*, 43(4):691–710.

Dunn, K. M., K. Jordan, and P. R. Croft, 2006. Characterizing the course of low back pain: a latent class analysis. *American Journal of Epidemiology*, 163(8):754–61. doi: 10.1093/aje/kwj100.

Dupre, M. E., A. N. Beck, and S. O. Meadows, 2009. Marital trajectories and mortality among US adults. *American Journal of Epidemiology*, 170(5):546–55.

Edin, K. and J. Reed, 2005. Why Don't They Just Get Married? Barriers to Marriage among the Disadvantaged. *The Future of Children*, 15(2):117–137.

Elder, G., 1994. Time, Human Agency, and Social Change: Perspectives on the Life Course. *Social Psychology Quarterly*, 57(1):4–15.

Elder, G. H., 1985. *Life course dynamics: trajectories and transitions, 1968-1980*. Ithaca, NY: Cornell Univ Press.

Elzinga, C., 2006. Sequence analysis: Metric representations of categorical time series. *Sociological Methods Research*, 38(3):463–481.

Elzinga, C. and A. Liefbroer, 2007. De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis. *European Journal of Population*, 23(3):225–250.

Elzinga, C., S. Rahmann, and H. Wang, 2008. Algorithms for subsequence combinatorics. *Theoretical Computer Science*, 409(3):394–404. doi:10.1016/j.tcs.2008.08.035.

Ermisch, J. and D. Pevalin, 2005. Early Motherhood and Later Partnerships. *Journal of Population Economics*, 18(3):469–489.

Espeland, M. and S. Handelman, 1989. Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics*, 45(2):587–599.

Evans, W., W. Oates, and R. Schwab, 1992. Measuring Peer Group Effects: A Study of Teenage Behavior. *The Journal of Political Economy*, 100(5):966–991.

Ferree, M., 1990. Beyond Separate Spheres: Feminism and Family Research. *Journal of Marriage and Family*, 52(4):866–884.

Fisher, R., 1955. Statistical Methods and Scientific Induction. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(1):69–78.

Francesconi, M., S. P. Jenkins, and T. Siedler, 2010. The effect of lone motherhood on the smoking behavior of young adults. *Health Economics*, in press.

Freedman, D. and D. Lane, 1983. A Nonstochastic Interpretation of Reported Significance Levels. *Journal of Business & Economic Statistics*, 1(4):292–298.

Furstenberg, F. F., 1976. *Unplanned parenthood: the social consequences of teenage childbearing*. New York: Free Press.

Furstenberg, F. F., 1998. When Will Teenage Childbearing Become a Problem? The Implications of Western Experience for Developing Countries. *Studies in Family Planning*, 29(2):246–253.

Furstenberg, F. F., 2005. Non-normative life course transitions: reflections on the significance of demographic events on lives. *Advances in Life Course Research*, 10:155–172.

Fussell, E., A. Gauthier, and A. Evans, 2007. Heterogeneity in the Transition to Adulthood: The Cases of Australia, Canada, and the United States. *European Journal of Population*.

Gabadinho, A., G. Ritschard, M. Studer, and N. S. Müller, 2009. Mining Sequence Data in R with TraMineR: A User's Guide. Technical report, Department of Econometrics and Laboratory of Demography, University of Geneva, Geneva. (TraMineR is on CRAN the Comprehensive R Archive Network).

Gauthier, J., E. Widmer, P. Bucher, and C. Notredame, 2009. How Much Does It Cost?: Optimization of Costs in Sequence Analysis of Social Science Data. *Sociological Methods Research*, 38(1):197–231.

George, L., 1993. Sociological perspectives on life transitions. *Annual Review of Sociology*, 19(1):353–373.

George, L. K., 2009. Conceptualizing Life course trajectories. In G. H. Elder and J. Z. Giele, editors, *The Craft of Life Course Research*. New York, NY: Guilford Press.

Giele, J. Z. and G. H. Elder, 1998. *Methods of life course research: qualitative and quantitative approaches*. Los Angeles, CA: SAGE Publications.

Goodman, L. A., 1974. Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models. *Biometrika*, 61:215–231.

Gove, W., 1972. The Relationship between Sex Roles, Marital Status, and Mental Illness. *Social Forces*, 51(1):34–44.

Gower, J. and W. Krzanowski, 1999. Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. *Journal of the Royal Statistical Society: series C*, 48(4):505–519.

Groff, E. R., D. Weisburd, and S.-M. Yang, 2010. Is it Important to Examine Crime Trends at a Local "Micro" Level?: A Longitudinal Analysis of Street to Street Variability in Crime Trajectories. *Journal of Quantitative Criminology*, 26(1):7–32. doi: 10.1007/s10940-009-9081-y.

Hadgu, A. and Y. Qu, 1998. A biomedical application of latent class models with random effects. *Journal of the Royal Statistical Society. Series C*, 47(4):603–616.

Hagenaars, J., 1988. Latent structure models with direct effects between indicators: local dependence models. *Sociological methods & research*, 16(3):379–405.

Hagenaars, J. A. and A. L. McCutcheon, 2002. *Applied latent class analysis*. Cambridge, UK: Cambridge University Press.

Halfon, N. and M. Hochstein, 2002. Life Course Health Development: An Integrated Framework for Developing Health, Policy, and Research. *The Milbank Quarterly*, 80(3):433–479.

Halpin, B., 2010. Optimal Matching Analysis and Life-Course Data: The Importance of Duration. *Sociological Methods Research*, 38(3):365–388. doi: 10.1177/0049124110363590.

Hamil-Luker, J. and A. M. O'Rand, 2007. Gender Differences in the Link Between Childhood Socioeconomic Conditions and Heart Attack Risk in Adulthood. *Demography*, 44(1):137–158.

Hamming, R., 1950. Error detecting and error correcting codes. *Bell System Technical Journal*, 26(2):147–160.

Harris, K. M., H. Lee, and F. Deleone, 2010. Marriage and Health in the Transition to Adulthood: Evidence for African Americans in the Add Health Study. *Journal of Family Issues*, forthcoming.

Harrison, W. J., B. M. Bewick, M. S. Gilthorpe, A. J. Hill, and R. M. West, 2009. Longitudinal latent class analysis of alcohol consumption. *Journal of Epidemiology & Community Health*, 63(Suppl 2):35–35. doi:10.1136/jech.2009.096719i.

Haviland, A., D. Nagin, and P. Rosenbaum, 2007. Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods*, 12(3):247.

Hayford, S. R., 2009. The Evolution of Fertility Expectations Over the Life Course. *Demography*, 46(4):765–783. doi:10.1353/dem.0.0073.

Hogan, D., 1978. The Variable Order of Events in the Life Course. *American Sociological Review*, 43(4):573–586.

Hogan, D. and N. Astone, 1986. The Transition to Adulthood. *Annual Review of Sociology*, 12:109–130.

Hollister, M., 2009. Is Optimal Matching Suboptimal? *Sociological Methods Research*, 38(2):235–264.

Hope, A., 1968. A Simplified Monte Carlo Significance Test Procedure. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(3):582–598.

Horwitz, A. and H. White, 1998. The Relationship of Cohabitation and Mental Health: A Study of a Young Adult Cohort. *Journal of Marriage and Family*, 60(2):505–514.

Jackson, P. and A. Berkowitz, 2005. The structure of the life course: Gender and racioethnic variation in the occurrence and sequencing of role transitions. *Advances in Life Course Research*, 9:55–90.

Johnston, D., 1995. Alternative Methods for the Quantitative Analysis of Panel Data in Family Research: Pooled Time-Series Models. *Journal of Marriage and Family*, 57(4):1065–1077.

Johnston, D., 2005. Two-wave panel analysis: Comparing statistical methods for studying the effects of transitions. *Journal of Marriage and Family*, 67(4):1061–1075.

Koball, H., E. Moiduddin, and J. Henderson, 2010. What Do We Know About the Link Between Marriage and Health? *Journal of Family Family*, forthcoming.

Kruskal, J., 1983. An Overview of Sequence Comparison: Time Warps, String Edits, and Macromolecules. *SIAM Review*, 25(2):201–237.

Lajunen, H.-R., A. Keski-Rahkonen, L. Pulkkinen, R. J. Rose, A. Rissanen, and J. Kaprio, 2009. Leisure activity patterns and their associations with overweight: a prospective study among adolescents. *Journal of Adolescent*, 32(5):1089–103. doi: 10.1016/j.adolescence.2009.03.006.

Lamb, K., G. Lee, and A. DeMaris, 2003. Union Formation and Depression: Selection and Relationship Effects. *Journal of Marriage and Family*, 65(4):953–962.

Landale, N., R. Schoen, and K. Daniels, 2010. Early Family Formation Among White, Black, and Mexican American Women. *Journal of Family Issues*, 31(4):445. doi: 10.1177/0192513X09342847.

Lazarsfeld, P. F. and N. W. Henry, 1968. *Latent Structure Analysis*. Boston, MA: Houghton Mifflin.

Lehrer, E., 1988. Determinants of marital instability: A Cox-regression model. *Applied Economics*, 20(2):195 – 210.

Lesnard, L., 2006. Optimal matching and social sciences. *Manuscript, Observatoire Sociologique du Changement (Sciences Po and CNRS), Paris.(http://laurent. lesnard. free. fr/)*.

Levenshtein, V. I., 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Dokl.*, 10:707–710.

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

Lichter, D., D. Graefe, and J. Brown, 2003. Is Marriage a Panacea? Union Formation among Economically Disadvantaged Unwed Mothers. *Social Problems*, 50(1):60–86.

Liefbroer, A. and E. Dourleijn, 2006. Unmarried cohabitation and union stability: testing the role of diffusion using data from 16 European countries. *Demography*, 43(2):203–221.

Lillard, L., M. Brien, and L. Waite, 1995. Premarital Cohabitation and Subsequent Marital Dissolution: A Matter of Self-Selection? *Demography*, 32(3):437–457.

Lillard, L. and L. Waite, 1993. A Joint Model of Marital Childbearing and Marital Disruption. *Demography*, 30(4):653–681.

Lin, H., B. Turnbull, C. McCulloch, and E. Slate, 2002. Latent Class Models for Joint Analysis of Longitudinal Biomarker and Event Process Data: Application to Longitudi-

nal Prostate-Specific Antigen Readings and Prostate Cancer. *Journal of the American Statistical Association*, 97(457):53–66.

Macmillan, R., 2005. The structure of the life course: classic issues and current controversies. *Advances in Life Course Research*, 9:3–24.

Macmillan, R. and S. R. Eliason, 2003. Characterizing the Life Course as Role Configurations and Pathways. In J. T. Mortimer and M. J. Shanahan, editors, *Handbook of the Life Course*, pages 529–554. New York, NY: Springer.

Manly, B. F. J., 1991. *Randomization and Monte Carlo methods in biology - Page 233*. London, UK: Chapman and Hall.

Martin, T. and L. Bumpass, 1989. Recent Trends in Marital Disruption. *Demography*, 26(1):37–51.

Mazzuco, S., 2009. Another look into the effect of premarital cohabitation on duration of marriage: an approach based on matching. *Journal of the Royal Statistical Society Series A*, 172(1):255–273.

McArdle, B. and M. Anderson, 2001. Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology*, 82(1):290–297.

McCutcheon, A. C., 1987. *Latent Class Analysis*. Beverly Hills, CA: Sage Publications.

McLanahan, S., 2007. Should government promote marriage? *Journal of Policy Analysis and Management*, 26(4):951–964.

McVicar, D. and M. Anyadike-Danes, 2002. Predicting Successful and Unsuccessful Transitions from School to Work by Using Sequence Methods. *Journal of the Royal Statistical Society. Series A*, 165(2):317–334.

Meadows, S., 2009. Family Structure and Fathers' Well-Being: Trajectories of Mental Health and Self-Rated Health. *Journal of Health and Social Behavior*, 50(2):115.

Miech, R. and M. Shanahan, 2000. Socioeconomic Status and Depression over the Life Course. *Journal of Health and Social Behavior*, 41(2):162–176.

Mirowsky, J., 2005. Age at First Birth, Health, and Mortality. *Journal of Health and Social Behavior*, 46(1):32–50.

Modell, F. F. Jr, and T. Hershberg, 1976. Social change and transitions to adulthood in historical perspective. *Journal of Family History*, 1:7–32.

Morey, L. and A. Agresti, 1984. The Measurement of Classification Agreement: An Adjustment to the Rand Statistic for Chance Agreement. *Educational and Psychological Measurement*, 44(1):33. doi:10.1177/0013164484441003.

Mouw, T., 2005. Sequences of Early Adult Transitions: A Look at Variability and Consequences. In R. Settersten, F. Furstenberg, and R. Rumbaut, editors, *On the Frontier of Adulthood: Theory, Research, and Public Policy*. Chicago, IL: University of Chicago Press.

Musick, K. and L. Bumpass, 2006. Cohabitation, marriage, and trajectories in well-being and relationships. *UC Los Angeles: California Center for Population Research. Retrieved from: http://www. escholarship. org/uc/item/34f1h2nt*.

Nagin, D. S. and R. Tremblay, 2005. Developmental Trajectory Groups: Fact Or A Useful Statistical Fiction? *Criminology*, 43(4):873–904.

Nock, S., 1981. Family Life-Cycle Transitions: Longitudinal Effects on Family Members. *Journal of Marriage and Family*, 43(3):703–714.

Nock, S., 2005. Marriage as a Public Issue. *The Future of Children*, 15(2):13–32. Marriage and Child Wellbeing.

O'Connell, M. and C. Rogers, 1984. Out-of-Wedlock Births, Premarital Pregnancies and their Effect on Family Formation and Dissolution. *Family Planning Perspectives*, 16(4):157–162.

Peters, A. and A. Liefbroer, 1997. Beyond Marital Status: Partner History and Well-Being in Old Age. *Journal of Marriage and Family*, 59(3):687–699.

Piccarreta, R. and F. C. Billari, 2007. Clustering work and family trajectories by using a divisive algorithm. *Journal of the Royal Statistical Society. Series A*, 170(4):1061 – 1078.

Pickles, A. and T. Croudace, 2010. Latent mixture models for multivariate and longitudinal outcomes. *Statistical Methods in Medical Research*, 19(3):271–89. doi: 10.1177/0962280209105016.

Radloff, L., 1977. The CES-D scale: A self report depression scale for research in the general population. *Applied Psychological Measurement*, 1:385–401.

Rand, W. M., 1971. Objective Criteria for the Evaluation of Clustering Methods. *Journal of American Statistical Association*, 66(336):846–850.

Reboussin, B., M. Miller, K. Lohman, and T. Have, 2002. Latent Class Models for Longitudinal Studies of the Elderly with Data Missing at Random. *Journal of the Royal Statistical Society. Series C*, 51(1):69–90.

Rendall, M., L. Clarke, H. Peters, N. Ranjit, and G. Verropoulou, 1999. Incomplete Reporting of Men's Fertility in the United States and Britain: A Research Note. *Demography*, 36(1):135–144.

Rindfuss, R., 1991. The young adult years: Diversity, structural change, and fertility. *Demography*, 28(4):493–512.

Rindfuss, R., C. Swicegood, and R. Rosenfeld, 1987. Disorder in the Life Course: How Common and Does It Matter? *American Sociological Review*, pages 785–801.

Roeder, K., K. Lynch, and D. S. Nagin, 1999. Modeling Uncertainty in Latent Class Membership: A Case Study in Criminology. *Journal of the American Statistical Association*, 94(447):766–767.

Saneinejad, S. and M. Roorda, 2009. Application of sequence alignment methods in clustering and analysis of routine weekly activity schedules. *Transportation Letters*, 1(3):197–211.

Savage, J. S. and L. L. Birch, 2010. Patterns of weight control strategies predict differences in women's 4-year weight gain. *Obesity (Silver Spring)*, 18(3):513–20. doi: 10.1038/oby.2009.265.

Schlich, R. and K. Axhausen, 2003. Habitual travel behaviour: Evidence from a six-week travel diary. *Transportation*, 30(1):13–36.

Schoen, R., N. Landale, and K. Daniels, 2007. Family Transitions in Young Adulthood. *Demography*, 44(4):807–820.

Schoen, R., N. Landale, K. Daniels, and Y. Cheng, 2009. Social Background Differences in Early Family Behavior. *Journal of Marriage and Family*, 71:384–395.

Schoenborn, C., 2004. Marital status and health: United States, 1999-2002. *Advance data*.

Seltzer, J., 2004. Cohabitation in the United States and Britain: Demography, Kinship, and the Future. *Journal of Marriage and Family*, 66(4):921–928.

Shanahan, M., 2000. Pathways to Adulthood in Changing Societies: Variability and Mechanisms in Life Course Perspective. *Annual Review of Sociology*, 26:667–692.

Shoval, N. and M. Isaacson, 2007. Sequence alignment as a method for human activity analysis in space and time. *Annals of the Association of American Geographers*, 97(2):282–297.

Soons, J. and M. Kalmijn, 2009. Is Marriage More Than Cohabitation? Well-Being Differences in 30 European Countries. *Journal of Marriage and Family*, 71(5):1141–1157.

Soons, J., A. Liefbroer, and M. Kalmijn, 2009. The Long-Term Consequences of Relationship Formation for Subjective Well-Being. *Journal of Marriage and Family*, 71(5):1254–1270.

South, S., 2001. The Geographic Context of Divorce: Do Neighborhoods Matter? *Journal of Marriage and Family*, 63(3):755–766.

Stovel, K. and M. Bolan, 2004. Residential trajectories: Using optimal alignment to reveal the structure of residential mobility. *Sociological methods & research*, 32:559–598.

Studer, M., G. Ritschard, A. Gabadinho, and N. S. Müller, 2010. Discrepancy Analysis of Complex Objects Using Dissimilarities. In H. B. et al., editor, *Advances in Knowledge Discovery and Management*, pages 3–19. Berlin: Springer-Verlag.

Teachman, J., 2002. Stability across cohorts in divorce risk factors. *Demography*, 39(2):331–351.

Teachman, J. and K. Crowder, 2002. Multilevel Models in Family Research: Some Conceptual and Methodological Issues. *Journal of Marriage and Family*, 64(2):280–294.

Tremblay, R. E., D. S. Nagin, J. R. Séguin, M. Zoccolillo, P. D. Zelazo, M. Boivin, D. Pérusse, and C. Japel, 2004. Physical aggression during early childhood: trajectories and predictors. *Pediatrics*, 114(1):e43–50.

Turney, K. and K. Harknett, 2010. Neighborhood Disadvantage, Residential Stability, and Perceptions of Instrumental Support Among New Mothers. *Journal of Family Issues*, 31(4):499. doi:10.1177/0192513X09347992.

Uebersax, J., 1999. Probit latent class analysis: Conditional independence and conditional dependence models. *Applied Psychological Measurement*, 23(4):283–297.

Vanhulsel, M., C. Beckx, D. Janssens, K. Vanhoof, and G. Wets, 2010. Measuring dissimilarity of geographically dispersed space–time paths. *Transportation*, forthcoming.

Vermunt, J., 2003. Multilevel latent class models. *Sociological Methodology*, pages 213–239.

Vermunt, J., 2008a. Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, 17(1):33.

Vermunt, J. K., 2008b. *Latent Class Models in Longitudinal Research*, chapter Handbook of Longitudinal Research: Design, Measurement, and Analysis. Burlington, MA: Elsevier.

Waite, L. J. and C. Bachrach, 2000. *The ties that bind: perspectives on marriage and cohabitation*. New York: Aldine de Gruyter.

Waldron, I., M. Hughes, and T. Brooks, 1996. Marriage protection and marriage selection–prospective evidence for reciprocal effects of marital status and health. *Social Science & Medicine*, 43(1):113–123.

Ward, J., 1963. Hierarchical grouping to optimize an objective function. *Journal of American Statistical Association*, (58):236–244.

Widmer, E. D. and G. Ritschard, 2009. The de-standardization of the life course: Are men and women equal?. *Advances in Life Course Research*, 14(1-2):28–39. doi: 10.1016/j.alcr.2009.04.001.

Wilson, C., 2001. Activity patterns of Canadian women: Application of ClustalG sequence alignment software. *Transportation Research Record*, 1777:55–67.

Wilson, C., 2006. Reliability of sequence-alignment analysis of social processes: Monte Carlo tests of ClustalG software. *Environment and Planning A*, 38(1):187–204.

Wilson, C., 2008. Activity patterns in space and time: calculating representative Hagerstrand trajectories. *Transportation*, 35:485–499.

Wood, R., B. Goesling, and S. Avellar, 2007. The effects of marriage on health: A synthesis of recent research evidence. Technical report, Department of Health and Human Services.

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

Wu, Z. and R. Hart, 2002. The Effects of Marital and Nonmarital Union Transition on Health. *Journal of Marriage and Family*, 64(2):420–432.

Zapala, M. and N. Schork, 2006. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proceedings*

of the National Academy of Sciences of the United States of America, 103(51):19430–19435.

A. R code

In this section, I present part of the R code used to implement the simulation study described in chapter 2.

The following R code is necessary to define the "sequence operators" described in chapter 2.

```
require("TraMineR")
require("poLCA")

#This is a function used to  extract a random subsequences
from the original subsequence
zac=function(l=n, r=15) {
p=rep(1/l,l)
repeat{ z1=rmulti(p); z2=rmulti(p);  if (abs(z1-z2)==r & abs(z1-z2)>1) break }
return(Z=list("z1"=z1, "z2"=z2)) }

#noise 1: inversion of a single sequence
noise1=function(seq=seq, level=.5){
n=length(seq)
seqx=seq
q=runif(1)
if (q<level ) {seq[seqx=="M"]="C"; seq[seqx=="C"]="M"}
return(seq) }
```

```
#noise2: mutation of a single sequence
noise2=function(s,level=0.5, alphabet){
l=length(alphabet)
q=runif(1)
p=rep(1/l,l)
a=rmulti(p)
if (q<level ) s=alphabet[a]
return(s) }

# cut: random truncation of a single sequence
cut=function(seq,level=0.5, k=15){
q=runif(1)
n=length(seq)
p=rep(1/k,k)
z1=k+rmulti(p)
if (q<level ) seq[z1:n]="NA"
return(seq)}

# noise4: postponement of a single sequence
noise4=function(seq=seq, level=.5){
n=length(seq)
for (g in 2:n){
q=runif(1)
if (q<level) seq[g]=seq[g-1]}
return(seq) }

#noise5: slicing of a single sequence
noise5=function(seq=seq, level=.5){
n=length(seq)
p=rep(1/(n-1),(n-1))
```

```
k=10
Z=zac(n, r=k)
z1=Z$z1
z2=Z$z2
a=seq[z1:z2]
Z=zac(n, r=k)
v1=Z$z1
v2=Z$z2
b=seq[v1:v2]
q=runif(1)
if (q<level){
seq[z1:z2]=b
seq[v1:v2]=a
}
return(seq)
}
```

The following code is used to implement the "sequence operators" on the entire dataset.

```
inversion=function(seq,l) seq2=t(apply(seq, 1, level=l, noise1))
slicing=function(seq,l) seq2=t(apply(seq, 1, level=l, noise5))
postponement=function(seq,l) seq2=t(apply(seq, 1, level=l, noise4))
mutation=function(seq,l) {if (is.seqe(seq)==FALSE)
A=seqdef(seq); alpha=alphabet(A)}
truncation=function(seq,l) seq2=t(apply(seq, 1, level=l, cut))
```

B. Additional tables and figures

In this appendix, I report the additional tables and figures used in the analysis presented in chapter 4. In particular, I present in table B.1 the costs setting used in the calculation of Optimal Matching distances in chapter 4. Tables B.2, B.3 and B.4 report the results of regression models using different indicators of *timing* and *quantum*. Last, Figure B.1 indicates the average time in each status by the membership to different typology of family trajectories.

Table B.1.: Substitution costs derived from data. Add-health, women of age 15-30

	C->	CP->	M->	MP->	S->	SP->
C->	0.0000	1.9907	1.9846	1.9998	1.9745	1.9999
CP->	1.9907	0.0000	2.0000	1.9865	1.9999	1.9653
M->	1.9846	2.0000	0.0000	1.9813	1.9942	2.0000
MP->	1.9998	1.9865	1.9813	0.0000	2.0000	1.9924
S->	1.9745	1.9999	1.9942	2.0000	0.0000	1.9982
SP->	1.9999	1.9653	2.0000	1.9924	1.9982	0.0000

Table B.2.: Regression estimates. Effects of timing indicators on health outcomes: age at first union

	(1) Poor Health	(2) Depression	(3) Smoking	(4) Drinking
Age at first union >25 (ref.)				
Age at first union <18	0.378***	0.625	3.175**	−0.121
	(0.0820)	(0.384)	(1.001)	(0.121)
Age at first union 19-20	0.238**	0.464	2.051*	−0.0702
	(0.0871)	(0.414)	(1.023)	(0.125)
Age at first union 21-25	0.126	0.176	0.979	−0.105
	(0.0805)	(0.387)	(0.935)	(0.117)
Age at wave I	−0.0570	−3.407	−12.75	0.978
	(0.860)	(5.332)	(13.43)	(0.783)
Age squared at wave I	0.000429	0.0916	0.320	−0.0300
	(0.0237)	(0.146)	(0.369)	(0.0215)
Living with bio-parents at wave I	−0.0540	−0.338	−1.119	−0.0129
	(0.0514)	(0.232)	(0.680)	(0.0765)
College educated parents	−0.198**	−0.620*	−3.338***	0.0563
	(0.0634)	(0.271)	(0.790)	(0.0984)
Hispanic	0.121	0.0115	−3.615***	−0.111
	(0.0834)	(0.378)	(0.806)	(0.0977)
Black	0.264***	0.349	−0.643	−0.277**
	(0.0627)	(0.314)	(0.806)	(0.0857)
Asian	0.247*	0.159	−1.778	−0.103
	(0.118)	(0.396)	(1.287)	(0.158)
Self-reported health at wave I	0.280***			
	(0.0291)			
Depression WI		0.275***		
		(0.0255)		
Smoking WI			0.509***	
			(0.0320)	
Drinking WI				0.176***
				(0.0279)
Constant	2.363	35.21	129.8	−7.137
	(7.800)	(48.68)	(122.2)	(7.155)
Observations	2168	2159	2155	2164
Adjusted R^2	0.149	0.113	0.303	0.054

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table B.3.: Regression estimates. Effects of timing indicators on health outcomes: age at first child

	(1) Poor Health	(2) Depression	(3) Smoking	(4) Drinking
Age at first children >25 (ref.)				
Age at first children <18	0.432***	0.771*	4.348***	0.223*
	(0.0839)	(0.382)	(1.060)	(0.104)
Age at first children 19-20	0.278**	−0.0794	2.867*	0.224
	(0.0980)	(0.411)	(1.232)	(0.121)
Age at first children 21-25	0.238**	0.469	1.538	0.0886
	(0.0808)	(0.330)	(0.869)	(0.0868)
Age at wave I	0.0200	−3.560	−0.921	1.279*
	(0.702)	(5.698)	(8.631)	(0.537)
Age squared at wave I	0.0000226	0.0971	0.00795	−0.0373*
	(0.0196)	(0.156)	(0.240)	(0.0148)
Living with bio-parents at wave I	−0.0291	−0.441	−1.997**	0.00160
	(0.0606)	(0.262)	(0.758)	(0.0727)
College educated parents	−0.215**	−0.677	−3.774***	−0.195*
	(0.0761)	(0.361)	(0.860)	(0.0796)
Hispanic	−0.0167	−0.380	−4.349***	0.0134
	(0.0967)	(0.425)	(0.945)	(0.114)
Black	0.0986	0.146	−2.838**	−0.251**
	(0.0705)	(0.379)	(0.931)	(0.0849)
Asian	0.353*	0.912	−2.027	−0.143
	(0.179)	(0.517)	(1.369)	(0.171)
Self-reported health at wave I	0.227***			
	(0.0335)			
Depression WI		0.234***		
		(0.0286)		
Smoking WI			0.535***	
			(0.0368)	
Drinking WI				0.131***
				(0.0286)
Constant	1.272	36.57	18.58	−10.51*
	(6.309)	(52.07)	(77.88)	(4.892)
Observations	1509	1504	1502	1505
Adjusted R^2	0.126	0.097	0.367	0.052

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure B.1.: Average time spent in each state by typology of trajectory

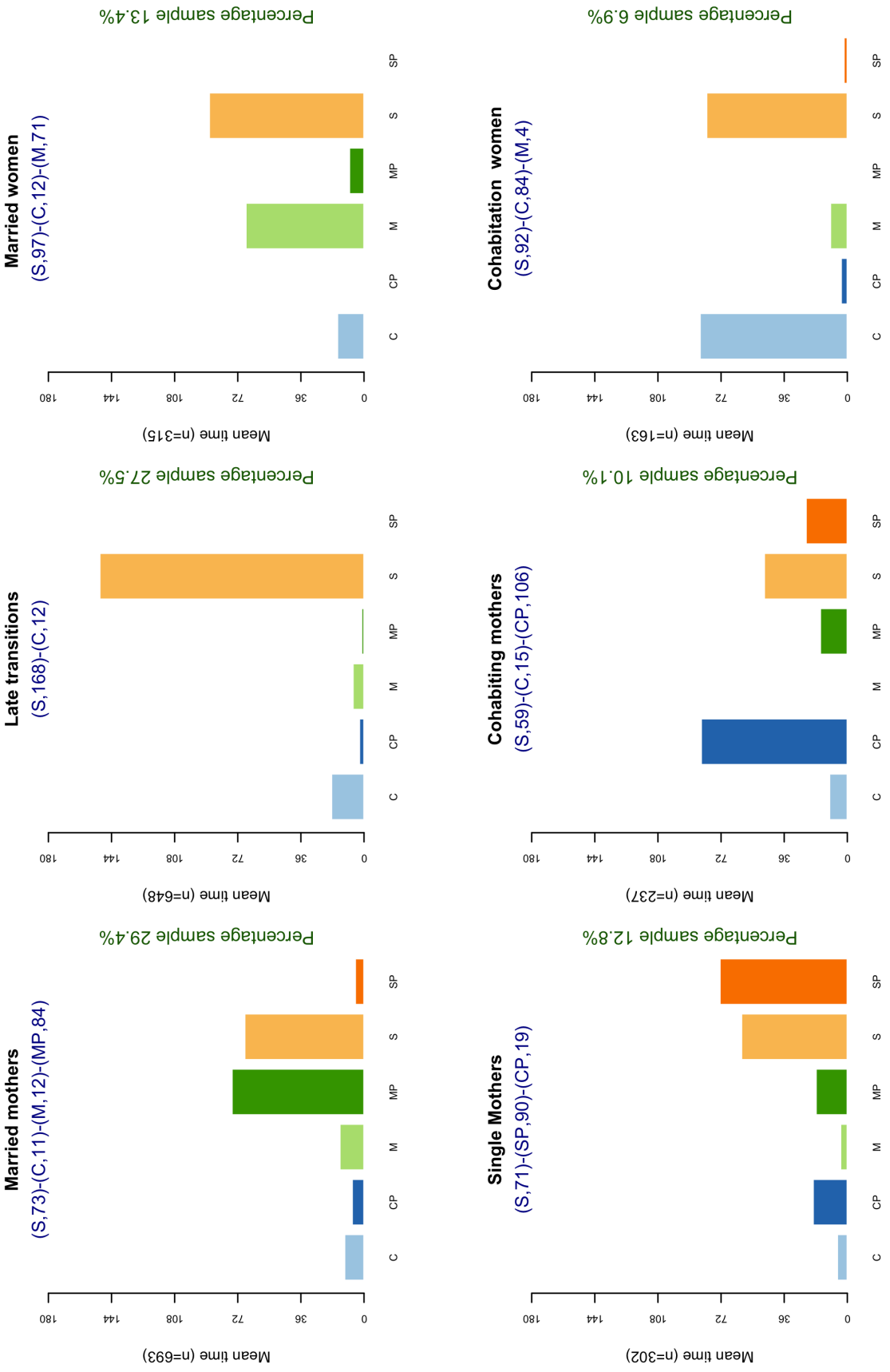


Table B.4.: Regression estimates. Effects of quantum indicators on health outcomes: sequences' turbulence.

	(1) Poor Health	(2) Depression	(3) Smoking	(4) Drinking
Turbulence	0.0217	0.0166	0.243*	0.000212
	(0.0114)	(0.0326)	(0.117)	(0.0135)
Age at wave I	0.329	−8.882*	−30.49*	0.562
	(1.477)	(4.495)	(15.03)	(1.108)
Age squared at wave I	−0.0105	0.242	0.806*	−0.0185
	(0.0406)	(0.124)	(0.410)	(0.0297)
Living with bio-parents at wave I	−0.0752	−0.106	−1.429*	−0.0103
	(0.0533)	(0.152)	(0.649)	(0.0722)
College educated parents	−0.274***	−0.492**	−3.553***	0.0694
	(0.0640)	(0.167)	(0.756)	(0.0914)
Hispanic	0.166	−0.121	−3.680***	−0.155
	(0.0912)	(0.238)	(0.749)	(0.0957)
Black	0.162**	0.417	−1.134	−0.283***
	(0.0620)	(0.221)	(0.783)	(0.0809)
Asian	0.190	−0.140	−1.850	−0.107
	(0.119)	(0.244)	(1.249)	(0.155)
Self-reported health at wave I	0.282***			
	(0.0295)			
CES-D scale at wave I		0.229***		
		(0.0254)		
Smoking WI			0.517***	
			(0.0316)	
Drinking WI				0.177***
				(0.0273)
Constant	−0.900	83.55*	292.0*	−3.461
	(13.43)	(40.79)	(137.7)	(10.32)
Observations	2254	2241	2236	2247
Adjusted R^2	0.128	0.086	0.305	0.060

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Barban Nicola

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.
Tel. +39 049 827 4111
e-mail: nicola@stat.unipd.it

Current Position

Since January 2007; (expected completion: July 2010)
PhD Student in Statistical Sciences. Department of Statistics, University of Padova.
Thesis title: Essays on sequence analysis for life course trajectories
Supervisor: Prof. Gianpiero Dalla Zuanna.

Research interests

- Life course analysis
- Immigrant assimilation
- Statistical methods for demographic research

Education

March 2004 2004-October 2006
Master degree in Statistics, Demographic and Social Sciences.
University of Padova, Faculty of Statistics
Title of dissertation: "The second generations of immigrants in Italy"
Supervisor: Prof. Gianpiero Dalla Zuanna.

September 2000-March 2004
Bachelor degree in Statistics and Management.
University of Padova, Faculty of Statistics
Supervisor: Prof. Lorenzo Bernardi.

Visiting periods

January-July 2009; January-April 2010
Visiting period at Population Studies Center.
University of Pennsylvania, Philadelphia PA, United States.
Supervisor: Francesco C. Billari and Frank J. Furstenberg

September-December 2009; April-July 2010
Visiting period at Dondena, Centre for Research on Social Dynamics.
Bocconi University, Milan, Italy.
Supervisor: Francesco Billari

April-July 2008
Visiting period at Population Studies and Training Center.
Brown University, RI, United States. Supervisor: Michael J. White

Further education

18-29 August 2009
Joint Summer School of the IUSSP and the Max Planck Institute for Demographic Research (MPIDR)
Summer Course: "Frontiers of Formal Demography"
Max Planck Institute, Rostock, Germany.
Organizers: Graziella Caselli (IUSSP), Heiner Maier (MPIDR).

August 2008
International Max Planck Research School for Demography.
Summer Course: "Immigrant Integration". (IMPRSD 153)
Lund University, Sweden. Instructor: Prof. Barry Chiswick (Chicago University)

March 2008
University of Tor Vergata, Rome, Italy.
"Tobit and selection models". Professor: Francis Vella (Georgetown University)

October 2007
University of Florence, Italy. Dipartment of Statistics "G. Parenti"
Course of the Italian Statistics Society (SIS): "Theory and practice of random effects models for multilevel and longitudinal data". Professors: Leonardo Grilli and Carla Rampichini

June 2006
Course of the Italian Statistics Society (SIS): 'Population and territory". Geographical information system (GIS); Multilevel Models; Spatial data; Simulation.

August 2005 - November 2005
Internship at: "Regional observatory on migration, Veneto Region" Italia Lavoro s.p.a.

February 2003 - July 2003
Erasmus project: Universidad de la Laguna, Tenerife (Spain)

Work experience

December 2007 - January 2008
Department of Statistical Sciences, University of Padua.
Supervisor of the second wave of ITAGEN2, CATI survey.

January 2006 - December 2006
Italia Lavoro s.p.a..
Junior researcher at "Regional observatory on migration, Veneto Region" (Statistical analysis on migration and demographic topics; research reports)

September 2001 - December 2001
Istat, Italian National Institute of Statistics
Census taker

Awards and Scholarship

January 2010
University of Pennsylvania. Visiting Doctoral Fellowship.
April 2006
Center for Voluntary Services, District of Padova. Award for dissertation on migration.
January 2003
University of Padova. Erasmus scholarship.

Computer skills

- OSX, Linux, Windows XP
- Microsoft Office (Word, Excel, Access, Powerpoint)
- Statistical packages (R, Stata, SAS, SPSS)
- Database (Access)
- Geographic Information System (MAPINFO)
- LaTeX

Language skills

Italian native. English, Spanish fluent.

Publications

Articles in journals

Barban, N. and G. Dalla Zuanna, (2010) "A portrait of immigrant childrens housing experiences in Italy". *Housing Studies*, 25(4): 559–584.

Chapters in books

Dalla Zuanna G. and N. Barban (2008) "Le seconde generazioni in Veneto", in *Studiare insieme, crescere insieme?* Franco Angeli, Milano. 2008

Dalla Zuanna G. and N. Barban (2007),"Giovani veneti, vecchi e nuovi" in *Immigrazione straniera in Veneto. Dati demografici, dinamiche del lavoro, inserimento sociale. Rapporto 2006.* Franco Angeli, Milano

Barban N. (2006), "La scuola" in *Immigrazione straniera in Veneto. Dati demografici, dinamiche del lavoro, inserimento sociale. Rapporto 2005.* Franco Angeli, Milano. May 2006

Working papers

"Immigrants childrens transition to secondary school in Italy" with Michael J. White *(submitted, available upon request)*

"Timing, quantum and sequencing in life course analysis. Are optimal matching and latent class analysis equivalent?"

"What Does Explain the Heterogeneity in Early Family Trajectories? A Non-Parametric Approach for Sequence Analysis"

Conferences

Barban N. and F. C. Billari "The analysis of life course trajectories: Sequence and latent class analyses compared" , *Statistical challenges in Lifecourse research. Royal Statistical Society*, Leeds, UK, 13/14 July 2010

Barban N. "What Does Explain the Heterogeneity in Early Family Trajectories? A Non-Parametric Approach for Sequence Analysis" (poster), *PAA Conference 2010*, Dallas, USA, 14/17 April 2010

Barban N. and M.J. White (2009), "The transition to secondary school of the second generation of immigrants in Italy" (poster), *IUSSP Conference 2009*, Marrakech, Morocco, 27 Sept./2 Oct. 2009

Barban N. and M.J. White (2009), "The transition to secondary school of the second generation of immigrants in Italy" (poster), *VID Conference on Education and Demography*, Vienna, Austria, 30 Nov./1 Dec. 2009

Barban N. and G. Dalla Zuanna (2008), "The homeownership of foreigners who live in Italy", *European Population Conference 2008*, Barcelona, 9-12 July 2008

Barban N. and G. Dalla Zuanna (2007) "The choice of buying a house for foreigners who live in Italy: Constrain or assimilation?", *Workshop: The Family and Residential Choice.* University of Amsterdam, 30 Aug./1 Sept. 2007

Other Interests

August 2004 voluntary service at the refugee camp of Cepin, Croatia
August 2006 voluntary service at the "Tuzlanska Amica" association, Tuzla, Bosnia-Erzegovina.
August 2007 voluntary service at the "Saint Martin" foundation, Nyahururu, Kenya.

References

Prof. Gianpiero Dalla Zuanna
Department of Statistical Sciences
University of Padova, 35121 Padova, Italy
Tel. +390498274190
e-mail: gpdz@stat.unipd.it

Prof. Francesco C. Billari
Department of Decision Sciences
Bocconi University,
Via Roentgen, 1, 20136 Milano, Italy
e-mail: francesco.billari@uni-bocconi.it

Prof. Frank F. Furstenberg
University of Pennsylvania,
3718 Locust Walk
277 McNeil/6299
Philadelphia, PA 19104
e-mail: fff@upenn.edu

Prof. Michael J. White
Brown University,
Population Studies and Training Center
68 Waterman Street,
Providence, RI 02912
e-mail: Michael_White@brown.edu