LIVES WORKING PAPER 2020 / 82

# THE LINK BETWEEN PREVIOUS LIFE TRAJECTORIES AND A LATER LIFE OUTCOME: A FEATURE SELECTION APPROACH

DANILO BOLANO

MATTHIAS STUDER

RESEARCH PAPER

http://dx.doi.org/10.12682/lives.2296-1658.2020.82

ISSN 2296-1658

The National Centres of Competence in Research (NCCR) are a research instrument of the Swiss National Science Foundation

**Auhtors**

Bolano, D. (1)

Studer, M. (2)

**Abstract**

Several studies have investigated the link between a previous trajectory and a given later-life outcome. Trajectories are complex objects. Identifying which aspects of the trajectories are relevant is of primary interest in terms both of prediction and testing specific theories. In this work, we propose an innovative approach based on data mining feature selection algorithms. The approach is in two steps. We start by automatically extracting several properties of the sequences. Using a life course approach, we focus here on features related to three key aspects of the life course: sequencing, timing and duration of life events. Then, in a second step, we use feature selection algorithms to identify the most relevant properties associated with the outcome. We discuss the use of two features selection approaches a random forest approach (Boruta) and a LASSO method (Stability Selection). We also discuss the inclusion of control variable such as socio-demographic characteristics of the respondent in this selection process. The proposed approach is illustrated through a study of the effects of family and work trajectories between age 20 and 40 on health and income conditions in midlife.

**Keywords**

**sequence analysis | variable selection | life course methodology | machine learning**

**Author's affiliation**

(1) NCCR LIVES, University of Lausanne

(2) LIVES Center and Institute of Demographics and Socioeconomics, Faculty of Social Sciences, University of Geneva

**Correspondence to**

danilo.bolano@unil.ch

# The Link Between Previous Life Trajectories and a Later Life Outcome: a Feature Selection Approach

Danilo Bolano[1] and Matthias Studer[2]

[1]*NCCR LIVES, University of Lausanne*
[2]*LIVES Center and Institute of Demographics and Socioeconomics, Faculty of Social Sciences, University of Geneva*

Corresponding author: danilo.bolano@unil.ch (Danilo Bolano)

**Abstract**

Several studies have investigated the link between a previous trajectory and a given later-life outcome. Trajectories are complex objects. Identifying which aspects of the trajectories are relevant is of primary interest in terms both of prediction and testing specific theories. In this work, we propose an innovative approach based on data mining feature selection algorithms. The approach is in two steps. We start by automatically extracting several properties of the sequences. Using a life course approach, we focus here on features related to three key aspects of the life course: sequencing, timing and duration of life events. Then, in a second step, we use feature selection algorithms to identify the most relevant properties associated with the outcome. We discuss the use of two features selection approaches a random forest approach (Boruta) and a LASSO method (Stability Selection). We also discuss the inclusion of control variable such as socio-demographic characteristics of the respondent in this selection process. The proposed approach is illustrated through a study of the effects of family and work trajectories between age 20 and 40 on health and income conditions in midlife.

## 1 Introduction

Many social sciences research questions are interested in the link between a previous trajectory and a given later-life outcome. The life course paradigm (e.g. Elder et al.,

2003) stresses the importance to situate any outcome within its individual temporal dynamics, and therefore to relate it with the previous trajectory. Moreover, life domains might be intertwined. Decisions and events happened in one domain may have a strong influence into another life domain. Therefore, one should take into account the unfolding of previous trajectories in other life domains as well.

Several studies have focused on the consequence of the life course on later life outcomes (Fasang, 2012). Social stratification researches have studied how and if previous life course patterns result in late life inequalities (Fasang, 2012; Gabriel et al., 2014). Social epidemiologists studied how previous professional, family or even housing tenure life-course shape aging, wellbeing and health condition at old age (e.g. McMunn et al., 2015; Sabbath et al., 2015; Hoven et al., 2017; Vanhoutte et al., 2017). Demographers analyzed how cohabitation trajectories influence leaving home (Rossignon et al., 2016). Criminologists studied how family formation events influence criminal careers (Zoutewelle-Terovan et al., 2012).

Two different strategies are generally used to study the link between previous trajectories and a given outcome (Rossignon et al., 2016). First, some authors include indicators of these trajectories in their models, such as whether the respondent experienced a given key event or the time spent in a given state (Gabriel et al., 2014; Blossfeld et al., 2007). However, the approach is limited, as previous trajectories might be complex resulting in the need to take many indicators into account, which might be relevant... or not (Rossignon et al., 2016). Furthermore, the current situation might result from complex interaction between past events that are difficult to identify. As a result, several authors use sequence analysis to operationalize the concept of past trajectories. This typically involves coding the trajectories as sequences, computing dissimilarities between them, creating a typology using cluster analysis and include it in their own analysis (e.g. Rossignon et al., 2016; Fasang, 2012; Gabriel et al., 2014; Hoven et al., 2017). The procedure has the advantage to take into account the timing, duration and sequencing of these trajectories, three key dimensions according to the life-course paradigm (Studer and Ritschard, 2016). Finally, the use of sequence analysis avoid making an a priori choice about the relevant dimensions of the previous trajectories probably resulting in the identification of more complex dynamics (Rossignon et al., 2016).

The use of sequence analysis raises several issues. First, the creation of a single typology aims to capture the timing, duration and sequencing of the previous trajectory altogether. However, it is often interesting to understand the more specific impact of each aspect taken separately (Studer and Ritschard, 2016). Indeed, each of these aspects might be linked to different sociological theories and models when studying the relationship between previous trajectory and outcome. For instance, the time spent in a state might capture the relationships between "time exposure" to unemployment and future health. The timing of a situation

or an event is thought to influence later life outcomes within the "critical period" model. According to this model, a situation such as unemployment only affect the outcome when it occurs within some specific age range, for instance after 50 years old. Finally, a link between specific sequencing and an outcome might reveal "mandatory steps" or the impact of specific "dynamics". For instance, many back-and-forth movements between employment and unemployment might have specific effect on future health.

Second, it might be difficult to give a clear interpretation of the association between previous trajectory typology and later outcome. Sequence analysis typologies are generally defined by an implicit set of rules (Studer, 2018). In other words, the typology works by grouping together similar trajectories. However, the exact properties of the sequences that distinguish the types are unknown. It might therefore be difficult to have an in-depth understanding of the links between the trajectory and a subsequent outcome.

Finally, any clustering technique works by simplifying the information. The aim is to ignore small variation in the sequences to build a few types. However, such simplification is made without taking the outcome into account. Therefore, the key variations in the sequences that explain the variation in the outcomes might very well have been "simplified" or ignored in the process. The procedure can therefore hinder the identification of some key properties of the sequences, therefore leading to wrong conclusions.

In this article, we propose an innovative methodology combining feature selection algorithms and sequence analysis tools to study the link between previous trajectory and later-life outcome. This combination aims to automatically identify the specific key dimensions of a previous trajectory that are linked with the outcome of interest. As such, it overcomes the indicator approach by taking many aspects of the previous trajectory into account, while providing information about the specific impact of timing, duration and sequencing.

The proposed methodology starts with the creation of a very broad and automatically defined set of indicators of the previous trajectory aiming to capture its timing, sequencing and duration dynamics. In a second step, it relies on machine learning feature selection algorithms to select the most relevant indicators, and therefore, the most important aspects of the trajectory for the domain under investigation. Two feature selection approaches are discussed: Boruta and LASSO. Boruta aims to select all relevant features associated with the outcome and can capture non-linear association and interaction effects between aspects of interest. LASSO aims to identify a parsimonious subset of features that best predict the outcome. We also discuss the inclusion of control variable such as age in this selection process to avoid the selection of confounders. The proposed methodology is illustrated through a study of the effect of previous family and professional trajectory on

self-rated health and income in midlife in Switzerland using longitudinal data from the Swiss Household Panel.

The remainder of this paper is organized as follows. In Section 2 we present the motivating examples. Data are discussed in Section 2.3. Section 3 discusses how to extract features from sequential data. In Section 4, we introduce our methodological approach discussing two ways of selecting the relevant features. In Section 5 we analyze the association between family and work histories, and health and wealth. The software used is briefly discussed in Section 6. Finally, Section 7 reports concluding remarks and discussion on future extensions.

# 2 Motivating Example: The Effect of Family and Working Life on Health and Income

In order to illustrate the proposed method and its usefulness, we rely on two sample applications. We focus on the consequences of professional and family trajectories before forty years old on self-rated health and household income in middle life. These two examples allow us to illustrate the use of the proposed methodology to study categorical (health in our example) as well as numerical (income) outcomes.

## 2.1 Working and Family Trajectories and Health

Drawing on life course and cumulative disadvantage literature (O'Rand, 2002; Dannesfer, 2003), many studies have focused on the link between late-life health inequalities and previous professional or family life trajectories.

The negative association between unemployment and physical and mental health has been explored in several studies (e.g. Paul and Moser, 2009; Falba et al., 2005; Laitinen et al., 2005; Strully, 2009). Most of them focused on the short-term effect of unemployment. Few exceptions include Clark et al. (2001) and Daly and Delaney (2013) that studied its long-term effects. Aside from the overall time spent in unemployment, having unstable employment histories is often though as a potential source of distress over a working life which might have a direct impact on health.

Family life trajectories can also play a role (Henretta, 2007; Kravadal et al., 2012; Williams et al., 2011; OFlaherty et al., 2016). Studies have found an association between parity and different types of later-life health conditions (e.g., Doblhammer, 2000; Hurt et al., 2006; Grundy and Kravdal, 2007; Grundy and Holt, 2000). The timing of childbearing is also relevant. Women who had children earlier in life had increased both metabolic (McMunn et al., 2015) and inflammatory markers (Lacey et al., 2015). Sironi (2018) found a non-linear effect of age at first child on health. Having a teenage pregnancy or having the first child after age 35 is associated with

an increased risk of chronic conditions later in life.

The marriage and its duration has also been found to be positively associated with health (Grundy and Holt, 2000; Henretta, 2010). The positive effect of marriage might be due to accumulation of social, emotional and economic support over a long period of time. However, early marriage might have negative effect on health (Grundy and Holt, 2000) probably due to higher risk of divorce later on and being disruptive for educational attainment.

Taking a broader perspective, Sabbath et al. (2015) and McKetta et al. (2018) studied the relationship between the unfolding of professional and family social roles over the life course and health outcomes. In these studies, role accumulation was linked with better health, while weak labor market relationship was generally associated with lower health in the US and UK.

## 2.2 Working Career, Family Life and Income

Losing a job is not only directly associated with income loss, but it might lead to deterioration of future labor market prospects (e.g. Heckman and Borjas, 1980; Arulampalam, 2008) and earnings due to a depreciation of human capital (Becker, 1993), underutilization and underdevelopment of skills. Social and working related stigma might be (see e.g. Gibbons and Katz, 1991). Employers might interpret career interruptions as a lack of commitment and ability with negative effects in terms of future wages and occupational advancements.

Due to different roles in the society expected for men and women, the effect of career interruption on earning differs by gender. Despite women experienced more often transitions in and out employment for childbearing, caring for children or older parents, the wage penalty for unemployment over the career is lower among women (Spivey, 2005).

The magnitude of the effects depends on the timing and duration of career interruption during the working life (Albrecht and Vromon, 1999; Beblo and Wolf, 2002; Spivey, 2005). Unsurprisingly, experiencing a longer period of unemployment (duration) is associated with lower re-employment's chances (Eriksson and Rooth, 2014). Youth unemployment has been found to have a scarring effect on later working life in terms of wages, working conditions and careers (Gregg, 2001).

According to Pohlig (2019), most studies have focused on the timing and duration of unemployment because of the use of simple indicators. He argues that we need a more detailed understanding of the consequence unemployment patterns and sequencing. This is one of the aims of the methodology proposed in this paper.

Having a child is often linked with career interruptions and lowers a woman's lifetime earnings (the so-called motherhood penalty) (Budig and England, 2001; Correll et al., 2007). The timing of career interruption due to childbearing matters. Studies have shown that motherhood gap is less important when childbearing is

postponed (Miller, 2011; Leung et al., 2016). These results are consistent with the human capital depreciation theory, where late career interruptions are thought to be less disruptive, as workers have already acquired the most important skills. Studies did not find a "fatherhood gap" but rather a "fatherhood premium" (Glauber, 2008).

Family unions have positive social and economic consequences (Ribar, 2004). Marriage has a positive effect on economic condition among men (marriage premium) (Hill, 1979; de Linde Leonard and Stanley, 2015). Its effect on women's earnings is, however, more complex to distinguish. Traditionally, childbearing follows marriage in subsequent years (typical sequencing of events). For these reasons, the potential positive effect of marriage on earning is often counterbalanced by the motherhood penalty (Budig and England, 2001).

## 2.3 Data

We use information on family and working life from the "Life History Calendar" questionnaire of the Swiss Household Panel (SHP) (Tillmann et al., 2016). This calendar collected detailed retrospective data on life trajectories in different life domains. In this work, we consider family history and working life between 20 and 40 years old on a sample of around 3,000 individuals (1626 women and 1378 men) aged 40 to 65 in 2013. For each individual we constructed yearly parallel trajectories of family and professional trajectories.

In order to compare the proposed methodology with the sequence analysis, we created a typology of family and professional trajectories taken separately. We used optimal matching with constant substitution costs to measure the distance between each pair of sequences. Then, we applied the partitioning around medoids (PAM) algorithm and retained the solution with the highest average silhouette width (Studer, 2013).

The professional trajectory allows distinguishing between: 1) Full-Time Work, 2) Part-Time Work, 3) Inactive or Unemployed (Non-Working), 4) Education. We add a fifth state in case of missing data. Missing information is quite rare with around 2% of missing states. The typology allows identifying seven groups of working trajectories (Figure 1) and five typologies of family history (Figure 2). From previous studies, we know that professional careers are highly gendered in Switzerland, resulting in different kinds of dynamics for women while men tend to follow a full-time breadwinner pattern (Levy et al., 2006; Widmer et al., 2003). This is confirmed in our sample, where 75% of men follow the full-time pattern while only 23% of women do so.

The cohabitation trajectories code the living arrangement until forty years old using five states: single, single with child(ren), with a partner, with a partner and child(ren) and other situations. We anticipate less structured cohabitation

Figure 1: Typology of Professional Trajectories

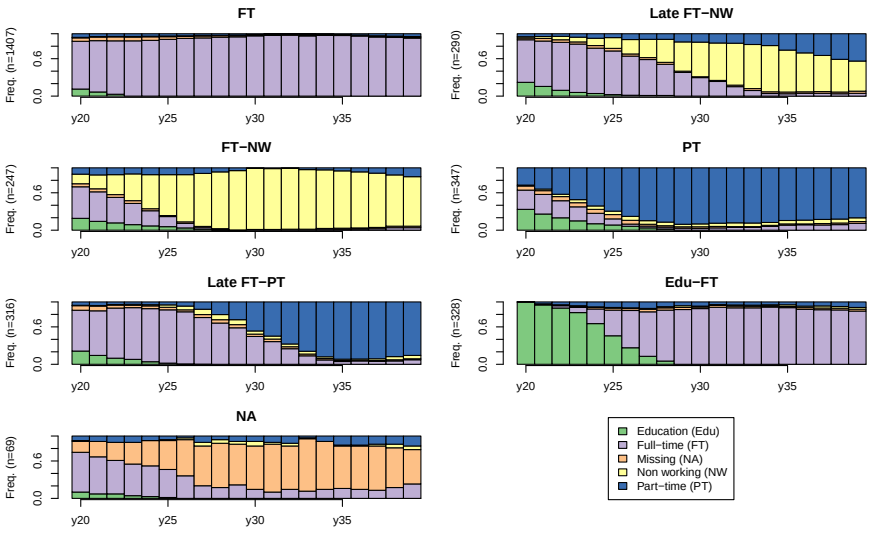
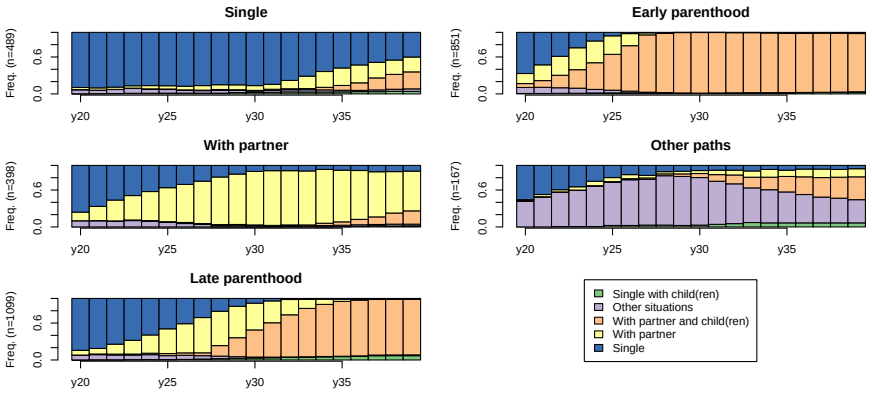



Figure 2: Typology of Cohabitation Trajectories




trajectories, but still centered around a few types (Levy et al., 2006). Figure 2 presents the five-group typology built using the same methodology as for professional trajectories. The most common pattern are early and late parenthood.

Using the prospective nature of the SHP data, we aim to analyze the effect of family life and working career on later-life self-rated health and later-life economic condition (income level). Self-rated health is measured as a binary variable, reporting “Very Well” or “Well” health versus other answers at the time of the

interview.[1] Income is measured using equalized disposable household income, since it takes family composition (household size) into account directly.

# 3 Trajectories' Features Extraction

The proposed methodology follows a three-step approach to identify relevant sequence properties. In the first step, we automatically extract several features of the trajectories. In the second step, we select the most relevant ones using data-mining feature selection algorithms. Finally, the remaining features are used as covariates in a traditional regression model to estimate the magnitude and the direction of their effects on the variable of interest.

## 3.1 Automatic Sequences Feature Extraction

The aim of the feature extraction is to build a set of indicators measuring the different aspect of the trajectories. In this article, we propose to build this set of indicators based on the three key aspects of life-course research identified by Studer and Ritschard (2016), namely sequencing, timing, and duration. We further propose some extensions of it.

The notion of sequencing relates to the order of the events or states within the trajectories. It regroups the quantum, i.e. which states or events occur, and the dynamics of the sequences coded through the succession of states (Billari et al., 2000). Both concepts can be thought to have an impact on later life outcomes. For instance, the succession and accumulation of roles is thought to have an impact on health in later life (McMunn et al., 2015). Traditional studies on the effect of life histories on later-life economic and health conditions might not be able to capture the importance of the order of events on the outcome since often rely on simple indicators such as experiencing the event or not (e.g., having or not the child), duration in a condition (e.g. duration of unemployment) and/or the time of events (e.g. age at first child).

Following Studer (2018), we rely on frequent subsequence mining to extract indicators of the sequencing of the states. This method aims to uncover frequent patterns of states out of a set of sequences (Studer et al., 2010; Agrawal and Srikant, 1995; Zaki, 2001).

A subsequence $s$ is defined as a subsequence of $x$ if all the states of $s$ occur in the same order as in $x$. For instance, the sequence $A - C$ is a subsequence of sequence $A - B - C$ because $A$ and $C$ occurs in the same order. In order to focus

[1] The self-rated health conditions are defined by the question "How do you feel right now?". Five possible answers were proposed: "not well at all", "not very well", "so, so" , "well", "very well".

on sequencing we look at the subsequence of the sequences of distinct successive states (i.e., without repeating a state to account for duration). A subsequence therefore captures either direct transition (two successive states) or medium term ordering. These two kinds of dynamics are interesting in many social sciences applications. For instance, when studying social mobility, direct transition can capture promotions, while medium term ordering should be able to measure job change, potentially with transitional unemployment spell, that results in upward mobility.

A subsequence might be of length one. Hence $A$ and $B$ are subsequences of $A - B$. These subsequences capturing the quantum of each situation along the trajectories. The notion of a subsequence is therefore able to capture the two aforementioned aspects of sequencing.

In this framework, a subsequence is said to be frequent if it is found in more than a predefined percentage of sequences, called the minimum support, that we typically set at 5%. Aside from a list of frequent subsequence, the analysis generates one variable per identified subsequence storing the presence or absence of the subsequence in each trajectory. It can therefore be used in any subsequent analysis, such as feature selection or a regression.

We extracted (see Table 1) 25 subsequences for the professional trajectories and 32 for the family one using in both cases a minimal support (i.e. frequency) of 5%. For instance, the pattern [Single $\rightarrow$ With partner $\rightarrow$ With partner and Kid] was followed by a little more than half of our sample. Aside from its frequency, we also have a dummy indicator for each individual whether this pattern occurs within his or her trajectory. Frequent subsequences are not mutually exclusive. Therefore, any individual following the [Single $\rightarrow$ With partner $\rightarrow$ With partner and Kid] also follows the [With partner $\rightarrow$ With partner and Kid] pattern as well.

The notion of timing relates to when a situation is experienced. The timing of events is crucial in many life-course researches. In the critical period model of life course epidemiology, events are often thought to have different consequences depending on when they occurred. Losing a job at the beginning or at the end of a professional career might have very different applications, but the same apply to the age at childbirth Lesnard (2010); Leung et al. (2016); Sironi (2018, e.g.).

Studer (2018) uses the situation at each age to capture the timing of the process. We take here a more parsimonious approach by computing the time spent in each state for each 5−years period, i.e. 20-24, 25-29, 30-34 and 35-39 years old. Table 2 presents the average of these indicators. This allows to quickly identify the general trend of the timing of these trajectories as well. Unsurprisingly, education mostly occurs at the beginning of the sequence (until age of 30), while living with kids and a partner occurs more frequently towards the end of the considered family life span when respondents had 30/35 years old.

Table 1: Frequent subsequences in family and professional trajectories and their associated frequencies.

| Work trajectories | | Family trajectories | |
|---|---|---|---|
| Subsequence | Perc. | Subsequence | Perc. |
| FT | 91.4% | Sin | 84.8% |
| PT | 45% | Part | 82.5% |
| FT → PT | 32.7% | Part/Kid | 76% |
| NW | 28.4% | Sin → Part | 69.8% |
| Edu | 26.2% | Part → Part/Kid | 65.2% |
| Edu → FT | 22.8% | Sin → Part/Kid | 62.3% |
| FT → NW | 22.1% | Sin → Part → Part/Kid | 54.8% |
| NA | 20% | Oth | 26.2% |
| FT → FT | 16.9% | Oth → Part | 16.8% |
| PT → FT | 15.4% | Oth → Part/Kid | 16.4% |
| FT → NA | 13.8% | Sin → Oth | 16.3% |
| NA → FT | 13.2% | Sin → Sin | 13.7% |
| Edu → PT | 12.5% | Oth → Part → Part/Kid | 12.2% |
| NW → PT | 12% | Part → Sin | 11.3% |
| PT → NW | 10.7% | Oth → Sin | 10.9% |
| PT → PT | 9.5% | Sin → Oth → Part | 9.6% |
| FT → NW → PT | 9.3% | Sin → Oth → Part/Kid | 9.4% |
| FT → NA → FT | 8.4% | Sin → Part → Sin | 8.9% |
| NA → PT | 8.2% | Sin/Kid | 8.5% |
| FT → PT → FT | 7.9% | Sin → Sin → Part | 8.4% |
| Edu → FT → PT | 7.7% | Part → Part | 8.2% |
| FT → PT → NW | 6.7% | Oth → Sin → Part | 7.2% |
| Edu → NW | 6.4% | Part → Sin → Part | 6.8% |
| NW → FT | 5.3% | Sin → Oth → Part → Part/Kid | 6.8% |
| FT → NA → PT | 5.1% | Sin → Part → Part | 6.4% |
| | | Sin → Oth → Sin | 6.1% |
| | | Part/Kid → Sin/Kid | 6.1% |
| | | Oth → Sin → Part/Kid | 5.9% |
| | | Sin → Sin → Part/Kid | 5.4% |
| | | Sin → Sin/Kid | 5.3% |
| | | Part → Sin/Kid | 5.3% |
| | | Sin → Part → Sin → Part | 5.3% |

Note: The states are abbreviated as follows for professional trajectories: FT (full time), PT (Part-time), NW (non-working), Edu (education), NA (missing value). For family trajectories: Sin (Single), Part (with partner), Part/Kid (with partner and kids), Oth (other situations), Sin/Kid (single with kids).

Finally, the notion of duration refers to the overall time spent in each state (Studer and Ritschard, 2016). It can be linked to the concept of "time exposure" to a given situation over the life course. As already discussed, duration of unemployment over the life course, length of marriage among others have strong effects on wellbeing. The overall time spent in each state in our sample is presented in Table 2 (last row). As observed also in Figure 1, respondents spent the majority of time being in a Full-Time working conditions (on average 12 years). For the family trajectories, the most common condition is having a partner and kids (Part/Kid state).

In some applications, it might be useful to distinguish further between the overall

Table 2: Average of Timing's and Duration's Indicators.

| | Work trajectories | | | | | Family trajectories | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Age group | Edu | FT | NA | NW | PT | Sin/Kid | Oth | Part/Kid | Part | Sin |
| 20-24 | 0.9 | 3.2 | 0.3 | 0.2 | 0.5 | 0.0 | 0.5 | 0.3 | 1.0 | 3.1 |
| 25-29 | 0.1 | 3.3 | 0.2 | 0.6 | 0.7 | 0.1 | 0.4 | 1.4 | 1.6 | 1.5 |
| 30-34 | 0.0 | 3.0 | 0.2 | 0.8 | 1.0 | 0.1 | 0.2 | 2.7 | 1.2 | 0.8 |
| 35-39 | 0.0 | 2.8 | 0.1 | 0.7 | 1.3 | 0.2 | 0.2 | 3.4 | 0.7 | 0.5 |
| Overall | 1.0 | 12.3 | 0.8 | 2.3 | 3.5 | 0.4 | 1.3 | 7.8 | 4.5 | 5.9 |

Note: The states are abbreviated as follows for professional trajectories: Edu (education), FT (full time), PT (Part-time), NW (non-working),NA (missing value). For family trajectories: Sin (Single), Part (with partner), Part/Kid (with partner and kids), Oth (other situations), Sin/Kid (single with kids).

time spent and typical spell duration. This can be included by adding the number of spells in each state for instance. When studying professional integration trajectories, this might be interesting to make the distinction between long unemployment spells and multiple short employment spells for instance. In our application, multiple spells of unemployment are rare and in general trajectories are quite stable over time with most individuals experience only one spell in each state during our observational period. We have therefore not included it in our analysis.

Aside from the key dimensions discussed above (Studer and Ritschard, 2016), we also added the complexity index (Gabadinho et al., 2010). It aims to capture the unpredictability and instability of the trajectories. The exact value of this index is not interpretable, but a higher value can be linked to a higher complexity of the trajectories.

## 3.2 Designing and Adapt the Set of Properties

Building on previous experiences, we argue that four key points should be taken into account while considering the features to extract. First, the properties should have a meaningful interpretation for the specific applications. Ideally, these indicators should be linked with relevant theoretical assumptions. This ensures that the results are interpretable. The features we have extracted here might not be relevant in all applications. We strongly encourage its adaptation to fit the research issue as much as possible. Some indicators could be removed, for instance if complexity is not of central interest, and others added. For instance, when studying professional integration trajectories it might be useful to add a specific indicator for the age of the individual at the time he/she lost his/her job.

Second, in order to have meaningful theoretical interpretation, the set of indicators should be as complete as possible. It should cover all the aspects of the trajectories that are relevant for the domain under study. In the end, we will

conclude on the effect of the previous trajectories taken as a whole, we therefore need to ensure that all relevant aspects were taken into account.

Third, each indicator should be interpretable on its own, even when the other indicators are not selected. Practically, a single concept should therefore not be coded using two (or more) variables. The procedure might end up selecting only one of the two making the interpretation difficult. For instance, the age at first birth should not be coded in two features like having the first baby and then, for those who experienced this event, the age at which is happened. If only one feature ends up selected, the remaining one will not represent the concept of timing of the transition to parenthood.

Finally, we should not have several indicators of the same underlying concept. Our set should be parsimonious for two reasons. It allows mitigating multiple testing issues and this is important even if most feature selection algorithms take it into account. It also typically leads to more stable results as feature selection methods might be affected by highly correlated features.

Parsimoniosity can be further achieved by removing duplicate and constant indicators. A constant indicator will never help in distinguishing the effect of the previous trajectory on a given outcome. For instance, in our example looking at trajectories up to age 40, nobody is in education in the last two age groups (30-35, 35-39) and the related indicators always equal 0. We therefore directly removed them from the list.

Some indicators might also be exact duplicates. In this case, it is better to remove them to avoid multicolinearity issues. These indicators can be identified by looking at the correlation matrix between the indicators. Several methods are available to do it automatically (e.g. Orestes Cerdeira et al., 2018; Jolliffe, 2002), even to exclude highly but not perfectly correlated items (Cadima and Jolliffe, 2001). We did not use these methods here, but they might be useful when a large number of features are automatically extracted. However, we removed the last age group of our timing features, as it can be directly derived using a linear combination of the other age groups and the overall time spent in each state.

Summarizing, we need to build the smallest but exhaustive possible set of theoretically relevant indicators, which are interpretable on their own.

In our empirical example, we extracted 45 indicators for the professional trajectory and 53 for the family one. We now turn to the selection of the relevant ones in studying the association between family and working histories, and income and health later in life.

In the rest of the paper, we will refer to all these indicators more generally as sequence features.

# 4 Feature Selection

The second step in our proposed approach is to identify relevant sequence features using data mining feature selection's methods. The aim of such methods is to automatically select a set of features that is statistically linked with the outcome variable.

There are two different kinds of feature selection methods. First, the "all-relevant" approach tries to identify all the features that are significantly linked with the outcome variable. Specialized methods are required to do so in order to take multiple testing into account. These methods are useful when we are interested in looking at an entire set of meaningful features (Degenhardt et al., 2017). However, as we will show, these methods often select a high number of possibly redundant features, making the exact interpretation of the effect of each feature difficult.

The second approach, often called "predictive", tries to directly select a relatively small set of features with the best predictive power, typically in a regression-like setting. However, by doing so, it discards relevant features that correlate with one or several of the selected features.

In this paper, we explore the use of both approaches and compare their results. For the "all-relevant" approach, we use Boruta (Kursa and Rudnicki, 2010), an extension of random forest, which was shown to feature among the best methods in this approach (Degenhardt et al., 2017). We rely on the well-known stability selection and LASSO regression (Meinshausen and Bühlmann, 2010) in the "predictive" approach. Before presenting these approaches, we discuss the issue of controlling for confounders in the selection process, as it applies to both methods.

## 4.1 Controlling for Confounders

When we study a quantitative relationship, we usually want to account for possible confounders. The same applies when studying the relationships between trajectories and income or self-rated health. We control for socio-demographic characteristics that might be correlated with health and economic condition. More specifically, we control for migration status (Swiss born or not), level of education (having at least tertiary education) and age to account for age difference at the time of the interview. Since we have cross-sectional data, age is also equivalent to the cohort, and the cohorts are strongly linked to life trajectories in Switzerland (e.g. Levy et al., 2006). We therefore want to control for age/cohorts to avoid selecting trajectories features because they measure in fact an age/cohort effect. Since professional and family histories are strongly gendered, we will perform the selection of the features stratified by gender.

In order to control for possible confounders, we use a two-step approach.

- Step 1. We start by estimating a regression model with the set of control variables on the outcome. We use a linear regression for income and binomial regression model for self-rated health.
- Step 2. We use the residuals of the model estimated in step 1 in the feature selection algorithm.

By doing so we ensure that our outcome is orthogonal to any confounder. Therefore, any selected feature should not be linked with a possible confounder but directly with the outcome. For the analysis of self-rated health, we used the residuals on the "deviance" scale resulting in a quantitative variable that can be analyzed with feature selection algorithms in linear models.

## 4.2 Boruta

The aim of Boruta is to identify all relevant features that are linked to an outcome variable (Kursa and Rudnicki, 2010). This is particularly interesting if we aim to understand in detail the relationship between trajectories measured through their features and the outcome of interest. Boruta featured among the best methods in a recent review of all-relevant feature selection methods, particularly in low-dimensional spaces, i.e. when there are fewer features than observations, as in our case (Degenhardt et al., 2017).

The all-relevant features approach faces two main challenges. First, one should take multiple testing into account. If we keep the usual 5%-significance threshold and have 100 independent features, we would end up with 5 false positives, i.e. features flagged as relevant even if there are independent. Second, as explained by Kursa and Rudnicki (2010), the lack of a direct relationship with the outcome does not mean that a feature is not important when considering other features. For this reason, we need to rely on so-called wrapper methods, that take multiple features into account at the same time.

Boruta is a wrapper method built around the random forest (RF) approach (for technical details please refer to Breiman, 2001). RF is commonly used in machine learning, and only recently it has been started using in social sciences too (Perry, 2013; Berk et al., 2016; Bruno et al., 2018), to rank variables according to their "importance". It requires relatively small model tuning is not too computational intensive and provides reliable performances.

Based on regression tree approach, RF runs the analysis over many sub-datasets made by randomly selecting features (ensemble algorithm). The outcome is a sort of model averaging over (possible thousands) regression trees. The advantages of RF are to reduce overfitting issues and mitigate instability of regression trees. RF can be used with a large number of variables, even if they are highly correlated. It identifies interactions and non-linear associations, and any type of variable (numeric,

categorical) can be used as input. The typical output of RF is a ranking of the importance of each variable included in the analysis.

The random forest algorithm does not take any decision about the selection of the features. It just ranks them. Boruta solves these issues by using a resampling technique to determine the statistical importance of the features. It distinguishes those who are "important" from those who are "unimportant".

The algorithm is based on "shadow features" made by shuffling the values of the original features. A RF approach is then applied and the importance of both the shuffled and the original features are computed. Boruta then compares the importance of the original and the shadow features and marks the original feature as important if its importance is constantly higher. Repeating this procedure iteratively and eliminating progressively irrelevant features, we end up with a robust set of features that have been frequently identified as important. This procedure allows Boruta to find all the relevant features, including those that are weakly but significantly associated with the outcome. For more details on Boruta, please refers to Kursa and Rudnicki (2010).

As mentioned above, Boruta relies on random forest and therefore on decision trees. For this reason, it is able to capture non-linear relationships between features and outcomes. It might also detect interaction effects, i.e. a feature that would have a different effect depending on the value of another feature. For instance, we might think that the effect of non-working on income at later age depends on whether a woman live with a partner or not. These properties and Boruta's ability to capture weak relationships tend to result in large set of important features.

Table 3 presents the features selected by Boruta for the income and the self-rated health focusing on women. The results for men are available in the appendix. The column "Imp." shows the standardized importance of each feature. A higher value indicates a more important feature. The column "Prop." measures the proportion of resampling in which the feature had a higher importance than the most important "shadow" feature. We highlighted in blue the feature that reached this importance at least 95% of the time as this threshold fully account for multiple testing. Boruta, as other feature selection methods, does not estimate the direction of the association between features and the outcome but it only shows their relative importance.

Focusing on income, no sequencing features were identified as important in the family life trajectory. The order of family life events has no direct or indirect effect on household income in midlife. However, the timing and the time spent in different living arrangement conditions (duration) matter. More specifically, the timing and duration of living with a partner, in union with kids (lower relatively importance), or single as the most relevant aspect of family trajectories (respectively named in the table as Part, Part/Kid and Sin) were identified as important. Surprisingly, living

Table 3: Features selected by Boruta for income and self-rated health for women.

Family Trajectories.

| Feature | Income | | Health | |
|---|---|---|---|---|
| | Imp. | Prop. | Imp. | Prop. |
| **Sequencing** | | | | |
| Part/Kid | | | 3.45 | 0.69 |
| Part → Part/Kid | | | 6.75 | 0.99 |
| Sin → Part/Kid | | | 4.61 | 0.92 |
| Sin → Part | | | 3.56 | 0.71 |
| Sin → Part → Part/Kid | | | 4.62 | 0.91 |
| **Timing** | | | | |
| 20-24 Part/Kid | | | 6.95 | 0.99 |
| 20-24 Part | 4.34 | 0.96 | 5.84 | 0.96 |
| 20-24 Sin | 3.06 | 0.73 | 5.62 | 0.97 |
| 25-29 Part/Kid | 3.76 | 0.90 | 6.85 | 0.99 |
| 25-29 Part | 5.12 | 0.98 | 4.99 | 0.93 |
| 25-29 Sin | 3.68 | 0.89 | 4.96 | 0.92 |
| 30-34 Sin/Kid | | | 3.34 | 0.64 |
| 30-34 Part/Kid | 2.83 | 0.67 | 5.49 | 0.97 |
| 30-34 Sin | 3.23 | 0.78 | 3.34 | 0.63 |
| **Duration** | | | | |
| Overall Sin/Kid | | | 5.37 | 0.96 |
| Overall Oth | | | 4.36 | 0.85 |
| Overall Part/Kid | 5.15 | 0.99 | 9.46 | 1.00 |
| Overall Part | 7.65 | 1.00 | 5.77 | 0.96 |
| Overall Sin | 5.76 | 0.99 | 6.74 | 0.99 |
| **Complexity** | | | | |
| C-index | | | 7.66 | 1.00 |

Work Trajectories.

| Feature | Income | | Health | |
|---|---|---|---|---|
| | Imp. | Prop. | Imp. | Prop. |
| **Sequencing** | | | | |
| Edu → FT | 3.07 | 0.75 | | |
| Edu → FT → PT | 3.39 | 0.86 | | |
| Edu → PT | 2.74 | 0.64 | | |
| FT | 3.43 | 0.83 | | |
| FT → PT | 4.16 | 0.95 | | |
| NA | | | 3.29 | 0.64 |
| NW → PT | | | 3.34 | 0.64 |
| **Timing** | | | | |
| 20-24 Edu | 4.52 | 0.97 | | |
| 20-24 FT | 5.55 | 0.99 | | |
| 20-24 NW | 3.63 | 0.87 | 7.03 | 0.99 |
| 20-24 PT | 3.17 | 0.77 | | |
| 25-29 FT | 4.24 | 0.95 | 4.61 | 0.89 |
| 25-29 NW | 5.27 | 0.99 | 5.09 | 0.95 |
| 25-29 PT | 3.72 | 0.88 | 3.26 | 0.61 |
| 30-34 FT | 3.87 | 0.92 | 5.65 | 0.98 |
| 30-34 NW | 2.89 | 0.68 | 4.13 | 0.84 |
| 30-34 PT | 3.80 | 0.93 | 4.53 | 0.88 |
| **Duration** | | | | |
| Overall Edu | 4.30 | 0.96 | | |
| Overall FT | 6.08 | 0.99 | 6.77 | 0.99 |
| Overall NW | 4.87 | 0.97 | 6.02 | 0.97 |
| Overall PT | 6.40 | 1.00 | 6.58 | 0.99 |

Note: Imp: standardized importance. Prop. proportion of resampling in which the feature had a higher importance than the most important. For the abbreviation of the states, please refer to Table 1.

single with children is not selected. Focusing on work trajectories, the time spent in the different types of employment status (education, full-time, non-working or part-time) were found important. Some sequencing features capturing education or the back-and-forth movements in the labor market were also selected as important, such as going back to part-time after being in a full-time position. These sequencing aspects are also found in the timing indicators. Despite the fact we control for education attainment in the first step of our selection procedure, the time spent in education at the beginning of the trajectory (i.e., staying in education after age of 20) has been selected as relevant.

Many features selected for income are also selected for health. For the latter, the timing and sequencing of the state living with a partner with kids seem to be of key importance, confirming the importance of the timing of child birth (McMunn et al., 2015). The timing and the overall time spent in full-time, part-time and non-working are also important.

The sequencing, timing and duration of working states seem to be less relevant for men than women with much fewer features selected (see Appendix A). This can be explained by the stability of male working trajectories that are characterized by long periods of full-time work. Several family features were selected. Interestingly, living with a partner and children is also associated with health, but at a later age. Separation (the subsequence Part → Sin) was been found to be associated with income and health among men but not women.

The strengths of Boruta are close to its limits. On the one hand, Boruta selects

a high number of features, including weak association, non-linear relationships and interaction effects among them. It therefore provides a "global" overview of sequence's features that are associated with the outcome of the interest. On the other hand, the high number of selected features makes the interpretation of the results less straightforward. For this reason, the method is particularly relevant if one aims to gain a global understanding of the features playing an important role in the relationship between previous trajectory and later life outcome.

## 4.3 LASSO and Stability Selection

LASSO regression model is a well-known approach for variable selection (Tibshirani, 1996; Hastie et al., 2015). Differently from Boruta, LASSO aims to find an efficient subset of features to predict the outcome variable, i.e. income or health in our case. LASSO uses a penalization method of non-zero coefficients in its estimation procedure. As a result, when the weight given to penalization is high, many coefficients are set to zero and are therefore unselected by the model. For further details on LASSO method, please refers to the seminal work of Tibshirani (1996).

Stability selection (SS) (Meinshausen and Bühlmann, 2010) extends LASSO by using resampling approach to overcome two of its common issues. First, LASSO does not provide any guidance on the number of features to keep. SS offers such guidance by looking at risk error, as in usual statistical reasoning. Second, when correlated variables are used in LASSO, the procedure tends to keep only one of them, and shrink toward zero the coefficient of the others, even if they were significant. SS milds this effect by using resampling techniques. This aspect is crucial while selection features from sequential data since indicators tend to be highly correlated.

In this article, we use an improvement of Stability Selection proposed by Shah and Samworth (2012) and further discussed by Hofner et al. (2015). It works by applying LASSO on random subsample of the data and look at "stable" features, i.e. feature that are often selected by the model.

When using SS, two parameters need to be set. Hofner et al. (2015) discuss in detail the choice of these parameters. First, we need to specify the minimum percentage of subsamples in which a variable was selected in order to be considered as stable. According to Hofner et al. (2015), any value between 0.5 and 1 can be used and should not have a relevant impact on the result. Following Meinshausen and Bühlmann (2010), we set to cut-off point at 0.6. Second, the per-family error rate (PFER) measures the expected number of false positive. It should be set at least to the significance threshold (typically $\alpha = 0.05$) to fully account for multiple testing, i.e. when each feature measures a completely different concept. Hofner et al. (2015) recommend taking a value in the range $[\alpha; m \cdot \alpha]$ where $m$ is the number of features. Setting the PFER to a value of $m \cdot \alpha$ would lead not to take

Table 4: Features selected by stability selection for income and self-rated health for women and men.

| | Women | | Men | |
|---|---|---|---|---|
| | Income | Health | Income | Health |
| 1 | Family: Overall Part* | Family: Part → Part/Kid* | Work: NW | Work: 30-34 NW* |
| 2 | Work: FT* | Family: 20-24 Part/Kid* | Work: Edu → FT | |
| 3 | Work: FT → PT* | | Work: Overall Edu** | |
| 4 | Work: NW | | Work: 30-34 FT* | |
| 5 | Work: Edu → FT* | | | |
| 6 | Work: 20-24 Edu* | | | |

Note: Features selected with $PFER = 1$ are highlighted in blue. For the abbreviation of the states, please refer to Table 1. A star indicates a feature selected by Boruta as well.

multiple testing into account.

Table 4 presents the features selected by the SS algorithm with $PFER = m \cdot \alpha$. As we expected, the set of features selected by stability selection is much smaller than the one estimated by Boruta. The number of features selected ranged from 1 to 6 according to the outcome and gender considered. With $PFER = \alpha$, no features were selected since the procedure would be too conservative as our features are highly correlated due to the stability of sequences. The features highlighted in blue were selected with $PFER = 1$, the default value in the R package "stabs".

According to SS and as expected, family trajectory features are mostly important for women. The overall time spent with a partner is linked with income, while the pattern and timing of children matter for self-rated health. No features from family history have been selected among men.

Apart from education, which matters for men and women, income among women is mostly linked to back-and-forth movement on the labor market as mostly indicators of sequencing have been kept. This includes the occurrence of a full time, or non-working spells as well as the full-time → part-time pattern. Among men, in addition to the time spent in education after age 20, the full-time employment after leaving education and having experienced a period of unemployment (NW) matter for income level in midlife. Not working between age of 30 and 34 is associated with later-life health as well.

## 4.4 Comparing Boruta and Stability Selections Features

In this section, we briefly compare the features selected by Boruta and SS. First of all, SS identifies a much smaller number of features that were also selected by Boruta, except the occurrence of not-working spell when studying income. However, timing and duration features of not-working were flagged as important for women. Features selected by both methods are indicated with a start in Table 4.

To further compare the results obtained by Boruta and Stability Selection methods, we first estimated the direct linear relationship between each Boruta's

Table 5: Comparison Features selected by Boruta and Stability Selection for income and self-rated health for women.

| Income | |
|---|---|
| Feature | |
| **Sequencing** | |
| *Work Trajectories* | |
| Edu → FT | * |
| Edu → FT → PT | A |
| Edu → PT | A |
| FT | * |
| FT → PT | * |
| **Timing** | |
| *Work Trajectories* | |
| 20-24 Edu | * |
| 20-24 FT | |
| 20-24 NW | A |
| 20-24 PT | |
| 25-29 FT | |
| 25-29 NW | S |
| 25-29 PT | |
| 30-34 FT | A |
| 30-34 NW | A |
| 30-34 PT | |
| *Family Trajectories* | |
| 20-24 Part | |
| 20-24 Sin | |
| 25-29 Part/Kid | A |
| 25-29 Part | A |
| 25-29 Sin | |
| 30-34 Part/Kid | A |
| 30-34 Sin | |
| **Duration** | |
| *Work Trajectories* | |
| Overall Edu | S |
| Overall FT | |
| Overall NW | A |
| Overall PT | |
| *Family Trajectories* | |
| Overall Part/Kid | A |
| Overall Part | * |
| Overall Sin | |

| Health | |
|---|---|
| Feature | |
| **Sequencing** | |
| *Work Trajectories* | |
| NA | |
| NW → PT | |
| *Family Trajectories* | |
| Part/Kid | |
| Part → Part/Kid | * |
| Sin → Part/Kid | A |
| Sin → Part | A |
| Sin → Part → Part/Kid | A |
| **Timing** | |
| *Work Trajectories* | |
| 20-24 NW | A |
| 25-29 FT | |
| 25-28 NW | A |
| 25-29 PT | |
| 30-34 FT | |
| 30-34 NW | |
| 30-34 PT | |
| *Family Trajectories* | |
| 20-24 Part/Kid | * |
| 20-24 Part | |
| 20-24 Sin | A |
| 25-29 Part/Kid | S |
| 25-29 Part | A |
| 25-29 Sin | |
| 30-34 Sin/Kid | A |
| 30-34 Part/Kid | |
| 30-34 Sin | |
| **Duration** | |
| *Work Trajectories* | |
| Overall FT | |
| Overall NW | |
| Overall PT | |
| *Family Trajectories* | |
| Overall Sin/Kid | A |
| Overall Oth | |
| Overall Part/Kid | |
| Overall Part | A |
| Overall Sin | |
| **Complexity** | |
| *Family Trajectories* | |
| C-index | A |

Note: For the abbreviation of the state, please refer to Table 1.
* = features selected by both methods; A = direct linear relationship is accounted for when controlling; S = direct linear relationship remains significant after controlling; empty = direct linear relationship is not significant in any case.

features and the outcome. We then estimated whether this direct linear relationship

is still significant when controlling for the SS features.

These results are presented in Table 5 for women and Table 8 for men as follows: "*" for features selected by both methods, "A" if the direct linear relationship identified for Boruta feature is Accounted for when controlling for SS features, "S" if it remains Significant after controlling, and nothing if the direct linear relationship is not significant.

The results allow us to illustrate the added value of each approach. First of all, several Boruta features have no direct linear relationship with the outcome (empty sign: 13 out of 29 for income and 18 out of 32 for health. For women). This can be explained by the ability of Boruta to capture non-linear and interaction effects. These interactions and non-linear relationships are neglected by linear approaches such as stability selection.

Most of the time, direct linear relationships found by Boruta are accounted for when controlling for SS features ("A" in the table). [2] In other words, only a few Boruta's features are still correlated with the outcome after controlling for SS features ("S" in the table).[3] The comparisons of the results for men yield to a very similar conclusion (see Appendix B).

Different reasons can drive this result. On the one side, SS aims to identify a small set of features highly correlated with the outcome using a linear model. So, it is strongly focused on finding the best predictive linear model. It is not surprising then that after controlling for SS features, some correlation found for Boruta features disappear. On the other side, it also means that the SS's features "represents" other sequences features that are correlated with the selected one. Multiple features might bring a similar information and SS will usually select only one of them. In other words, SS features might hinder other correlated features. We should therefore take that into account in our interpretation of the results.

# 5 Regression Models

Feature selection algorithms allow us to identify the most relevant properties, but not to understand their effect. As a final step, we run a regression model including the SS features to interpret the relationships. We do not consider Boruta features, as their number and correlation would lead to multicolinearity issues.

Table 6 presents several logistic regression on health for women. The base model only includes our control variables. As expected, the chances to have a good self-rated health decrease with age and tends to be higher for those born in Switzerland. In the second model, the sequence analysis typologies of professional

[2] 10 features out of 16 for income and 11 out of 14 for self-rated health.

[3] Namely, for income: not-working between ages 25-29; overall duration of education after age 20; for health: having a partner and children at ages 25-29.

Table 6: Logistic regressions on health for women.

| | Base | Seq. An. | Stab. Sel. | Complete |
|---|---|---|---|---|
| Intercept | $2.74^{***}$ | $2.33^{***}$ | $2.32^{***}$ | $2.19^{***}$ |
| | (0.54) | (0.59) | (0.55) | (0.59) |
| Age | $-0.03^{**}$ | $-0.02^{*}$ | $-0.02^{*}$ | $-0.02$ |
| | (0.01) | (0.01) | (0.01) | (0.01) |
| Tertiary educ. | $-0.02$ | $-0.04$ | $-0.09$ | $-0.09$ |
| | (0.14) | (0.14) | (0.14) | (0.15) |
| Born in CH | $0.31^{*}$ | $0.34^{*}$ | $0.30^{*}$ | $0.34^{*}$ |
| | (0.14) | (0.14) | (0.14) | (0.15) |
| Family Life | | | | |
| Part → Part/Kid | | | $0.49^{***}$ | $0.61^{**}$ |
| | | | (0.14) | (0.20) |
| 20-24 Part/Kid | | | $-0.20^{***}$ | $-0.15^{*}$ |
| | | | (0.05) | (0.06) |
| Professional Cluster (ref=Full-time) | | | | |
| Late FT-NW | | $-0.08$ | | $-0.26$ |
| | | (0.24) | | (0.25) |
| FT-NW | | $-0.20$ | | $-0.34$ |
| | | (0.25) | | (0.26) |
| PT | | $-0.01$ | | $-0.11$ |
| | | (0.24) | | (0.24) |
| Late FT-PT | | $-0.36$ | | $-0.51^{*}$ |
| | | (0.23) | | (0.23) |
| Edu-FT | | $-0.07$ | | $-0.15$ |
| | | (0.32) | | (0.32) |
| NA | | 0.24 | | 0.20 |
| | | (0.44) | | (0.44) |
| Family life Cluster (ref=Single) | | | | |
| Early parenthood | | $-0.07$ | | $-0.15$ |
| | | (0.23) | | (0.29) |
| With partner | | 0.22 | | 0.22 |
| | | (0.27) | | (0.27) |
| Other paths | | 0.32 | | 0.32 |
| | | (0.41) | | (0.41) |
| Late parenthood | | $0.60^{*}$ | | 0.25 |
| | | (0.24) | | (0.28) |
| AIC | 1328.57 | 1330.39 | 1303.28 | 1311.84 |
| BIC | 1349.63 | 1404.11 | 1334.87 | 1396.09 |
| Log Likelihood | $-660.28$ | $-651.20$ | $-645.64$ | $-639.92$ |
| $\Delta\chi^2$ (vs Base) | | 18.18 | $29.29^{***}$ | $40.73^{***}$ |
| $\Delta\chi^2$ (vs Complete) | $5.94^{***}$ | $22.55^{***}$ | 11.44 | |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Note: For the abbreviation of the state, please refer to Table 1. Cluster of Sequence Analysis are reported in Figure 1 and 2.

and family life trajectories were added. Consistently with the literature, "late parenthood" type is associated with higher health than those in the "single" or "early parenthood" clusters. No significant effect is found for working career typology. More globally, a likelihood ratio test against the base model shows no significant overall improvement.

The third model includes the two SS features and shows a significant improvement over the base model ($\Delta\chi^2 = 29.29$; $p < 0.001$). The sequencing feature "Part → Part/Kid" (living with children and partner after a living with a partner episode) is associated with higher health. The feature "20-24 Part/Kid" (time spent in living with partner and children between 20 and 24 years old) is associated with lower self-rated health. In order to correctly interpret the proper effect of each property, the reference it should be clearly identified. In our case, it is a woman who did not have a child early in life, and not lived with a partner and a child after living with a partner. In our sample, this is mostly women without

children. Summarizing, we found the lowest self-rated health for early mothers, followed by women without children and finally late parenthood. All these results are consistent with previous studies on the link between family trajectories and midlife health conditions showing the negative effect of having children early in life (McMunn et al., 2015).

The last model includes the typologies and the SS features to provide a benchmark. It shows that adding SA typologies on top of SS features does not significantly improve the model (likelihood ratio test $\Delta\chi^2 = 11.44$). In this empirical example, the SS features capture all the relevant information provided by the SA typologies. On the contrary, comparing the last and the SA model, we can conclude that adding SS features to the SA model significantly increases the model fit (likelihood ratio test $\Delta\chi^2 = 22.55^{***}$). SS features therefore provide a more precise information as these features might be shared by individuals classified in different types. It also provides a much clearer interpretation.

Table 7 presents linear regression models of household income following the same modeling strategy as for health. The SA model (second column) highlights two significant and positive effects on income: living with a partner without child(ren) and the role of the time spent in education (typology labeled "Edu — Full-time" type). Unsurprisingly, couple without children is linked with higher household equivalized income as well as further education. The SS model provides a more precise interpretation. Aside from living with a partner and education, it shows the importance of working full-time, the full-time $\rightarrow$ part-time pattern (both positively associated with later-life income), experiencing nonworking spells (negative impact on income in midlife). The effect of the Edu $\rightarrow$ FT pattern is not significant, probably because of multicolinearity ($VIF = 2.9$). Here again, this model significantly improves the base model.

As before, a proper interpretation of these relationships between work-related features and household income requires us to identify the reference individual: a woman who did not experience any full-time or non-working spells and without time spent in education between 20 and 24 years old. These are predominantly women experiencing a full part-time trajectory.

We then observe a significant effect of experiencing a full-time spell and a negative effect of non-working spells. These results were expected. Surprisingly, the "FT $\rightarrow$ PT" pattern (full-time followed by a part-time spell) is associated with higher household income. This could be explained by a reverse causality, female workers who decided to reduce their labor force participation are also those who live already in a wealthy family and can afford it.

Here again, the complete model shows us that all the effects identified by SA typologies are accounted for by the SS features (i.e. they are not significantly associated with the outcome when including SS features). No significant improvement

Table 7: Linear Regression on Equivalized Household Income for women.

| | Base | Seq. An. | Stab. Sel. | Complete |
|---|---|---|---|---|
| Intercept | 33595.51*** | 23152.00** | 9352.45 | 7511.66 |
| | (8143.71) | (8785.52) | (8909.74) | (9650.40) |
| Age | 472.33** | 613.10*** | 683.49*** | 675.52*** |
| | (149.75) | (151.23) | (147.59) | (149.77) |
| Tertiary educ. | 14295.51*** | 12412.24*** | 10511.73*** | 10687.09*** |
| | (2148.27) | (2162.95) | (2158.73) | (2169.23) |
| Born in CH | 4023.59 | 4567.69* | 5220.70* | 5185.78* |
| | (2166.80) | (2157.51) | (2142.27) | (2164.30) |
| Family Life | | | | |
| Overall Part | | | 959.41*** | 922.78* |
| | | | (232.40) | (399.32) |
| Working Life | | | | |
| FT | | | 7522.80* | 8091.90 |
| | | | (3666.16) | (4234.45) |
| FT → PT | | | 5977.57** | 7003.72* |
| | | | (2280.57) | (2774.28) |
| NW | | | −6214.28** | −7008.99* |
| | | | (2143.38) | (3121.32) |
| Edu → FT | | | 1385.17 | 1130.34 |
| | | | (4151.16) | (4201.92) |
| 20-24 Edu | | | 3972.86*** | 3678.14** |
| | | | (1066.36) | (1161.58) |
| Professional Cluster (ref=Full-time) | | | | |
| Late FT-NW | | −2188.74 | | −346.42 |
| | | (3562.81) | | (4501.34) |
| FT-NW | | −7193.93 | | 125.31 |
| | | (3840.47) | | (4760.45) |
| PT | | 2390.33 | | 1559.38 |
| | | (3494.07) | | (4051.02) |
| Late FT-PT | | 2229.19 | | −3952.89 |
| | | (3482.17) | | (4001.61) |
| Edu-FT | | 18376.90*** | | 3450.20 |
| | | (4722.45) | | (5862.58) |
| NA | | −305.77 | | −956.49 |
| | | (6194.77) | | (6164.14) |
| Family life Cluster (ref=Single) | | | | |
| Early parenthood | | 979.11 | | 3163.78 |
| | | (3701.11) | | (3716.98) |
| With partner | | 12117.93** | | 2602.80 |
| | | (4159.02) | | (5977.35) |
| Other paths | | −3917.41 | | −3079.26 |
| | | (5721.13) | | (5677.09) |
| Late parenthood | | 4078.11 | | 2862.18 |
| | | (3640.47) | | (3698.68) |
| $R^2$ | 0.03 | 0.06 | 0.09 | 0.09 |
| $\Delta F$ (vs Base) | | 4.72*** | 15.06*** | 5.94*** |
| $\Delta F$ (vs Complete) | 5.94*** | 7.74*** | 0.49 | |
| BIC | 35969.09 | 35995.23 | 35924.61 | 35992.73 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Note: For the abbreviation of the state, please refer to Table 1. Cluster of Sequence Analysis are reported in Figure 1 and 2.

is found by adding the SA typologies to the SS model. Unsurprisingly, here again the SS approach provides the most parsimonious model.

We do not discuss in detail the model for men as they lead to similar conclusion. In all cases, SS features were able to detect the association found with the SA typologies. Furthermore, it also makes it possible to interpret more precisely the effect of the previous trajectory on the outcome of interest. Finally, in some models, additional interpretations were made possible by identifying features that were not taken into account in the typology.

# 6 Running the Analysis in R

We briefly present here how to run the proposed methodology in `R`. Nevertheless, feature selection approaches are available in other software too.

The automatic extraction of features is available using the `seqpropclust` function available in the `WeightedCluster` package (Studer, 2013, 2018). Aside from the state sequence object `myseq`, one needs to specify the `properties` to be extracted, and the `prop.only=TRUE` arguments to retrieve the list of extracted features. The function can extract other properties as well, see Studer (2018) for more information.

```
library(WeightedCluster)
## Extracting features
features <- seqpropclust(myseq,
            properties=c("pattern","agerange", "duration"),
            prop.only = TRUE)


## Controlling for confounders
regconfond <- glm(outcome~confounders)


## Running Boruta
library(Boruta)
Boruta(residuals(regconfond)~features)


## Running stability selection
library(stabs)
stabsel(features, residuals(regconfond), cutoff=0.6, PFER=1)
```

Once the features are extracted, the Boruta can be run using the `Boruta` function from the package of the same name (Kursa and Rudnicki, 2010). The `outcome` variable and the sequence `features` are specified through the usual `R` formula interface. Additional parameters of interest include `maxRuns` and `num.trees` which can be both increased to get more stable results.

Stability selection using LASSO regressions is available in the `stabs` package (Hofner and Hothorn, 2017) and the `stabsel` function. The `features` and `outcome` should be specified. Threshold for stability can be set using the `cutoff` argument, and the acceptable error through the `PFER` argument.

# 7 Conclusion

In this article, we proposed a new methodological approach to study the link between a previous trajectory and a given outcome. It starts by automatically extracting properties of trajectories coded as sequences. We proposed indicators to capture meaningful aspects of trajectories in a life-course perspective, namely duration, complexity, sequencing, and timing. Although we discussed an automatic feature extraction procedure, we also emphasized the need to adapt it for specific applications, by removing or adding indicators.

In a second step, relevant features are selected among the previously defined set of properties. We discussed two approaches. First, Boruta is an all relevant feature selection procedure able to capture non-linear and interaction effects. Second, stability selection using LASSO regression identifies a relevant subset of features to be included in subsequent analysis. These two approaches are complementary, as the first one allows to take a broader look on relevant trajectories' properties, while the second allows identifying a relevant subset for future uses. While presenting this framework, we also discussed how to consider possible confounders based on residuals.

Finally, the effects of relevant features can be interpreted using regression modeling. In our sample applications, we saw that the approaches provided more precise results and interpretation than the usual sequence analysis approaches. Our illustrative examples on the effect of family and working lives on later-life economic and health condition further showed that the proposed approach provides more precise results and interpretation than commonly used methods such as sequence analysis typologies or user-defined indicators.

The proposed approach is complementary to the more traditional SA typology. Both approaches aim to summarize the complexity of life trajectories, but they have different aims. The goal of sequence typologies is to provide a descriptive "global" view on trajectories, by identifying similar patterns. The feature selection approach proposed here aims to identify the specific characteristics of the trajectories that are significantly linked to a later-life outcome. Depending on the research question, one or the other approaches could be more useful.

While the proposed approach is promising for many applications, we see rooms for future developments. First, Boruta results showed us interaction and non-linear effects between trajectory indicators that should be taken into account. The resulting set of selected features is relatively large. Further work is needed to include them in the regression model afterward. This would allow us to study the combined effect of the intertwinement of different life domains, one of the core principles of the life-course paradigm.

Many feature selection algorithms were developed in the data-mining literature. We used here two approaches that are well recognized. However, other approaches

might be useful in life-course research too. Among others, the group-LASSO approach, a method that takes prior information on the grouping of the features into account (Yuan and Lin, 2006), might be interesting to consider life-domain interactions.

In this empirical example, the professional and family trajectories were considered separately, without accounting directly for the interaction between them, even if Boruta might have identified some of them. Future studies might look more specifically on life domains interactions using, for instance, multichannel techniques.

Finally, future works should focus on missing data to better understand how to properly treat them. In this work, missing states was infrequent and was mostly made of short episode. The results are therefore probably not strongly affected by missing data.

The proposed approach can also be used as starting point for further complex analysis. For instance, for causal inference, the features selected by Boruta could be used as matching variables in a propensity score matching framework if we want to match individuals according to their previous trajectories. Further studies are needed to compare this to existing approaches that use the (dis)similarity between previous trajectories as matching criteria (e.g. Barban et al., 2017).

# References


Agrawal, Rakesh and Ramakrishnan Srikant. 1995. "Mining Sequential Patterns." In *Proceedings of the International Conference on Data Engeneering (ICDE), Taipei, Taiwan*, edited by Philip S. Yu and Arbee L. P. Chen, pp. 487–499. IEEE Computer Society.

Albrecht, James W and Susan Vromon. 1999. "Unemployment Finance Compansetion and Efficiency Wages." *Journal of Labor Economics* 17:141–167.

Arulampalam, Wiji. 2008. "Is Unemployment Really Scarring? Effects of Unemployment Experiences on Wages." *The Economic Journal* 111:586–606.

Barban, Nicola, Xavier de Luna, E Lundholm, I Svensson, and Francesco C. Billari. 2017. "Causal Effects of the Timing of Life-course Events: Age at Retirement and Subsequent Health." *Sociological Methods & Research* pp. 1–34.

Beblo, Miriam and Elke Wolf. 2002. "How much does a year off cost? Estimating the wage effects of employment breaks and part-time periods." *Brussels Economic Review* 45:191–217.

Becker, Gary S. 1993. *Human Capital: A theoretical and Empirical Analysis with Special Reference in Education*. The University of Chicago Press. 3rd Edition.

Berk, Richard A, Susan B Sorenson, and Geoffrey Barnes. 2016. "Forecasting domestic violence: a machine learning approach to help inform arraignment decisions." *J. Empir. Legal Stud* 13:94–115.

Billari, Francesco C., Johannes Fürnkranz, and Alexia Prskawetz. 2000. "Timing, sequencing, and quantum of life course events: a machine learning approach." Working Paper 010, Max Planck Institute for Demographic Research, Rostock.

Blossfeld, Hans-Peter, Katrin Golsch, and Götz Rohwer. 2007. *Event History Analysis with Stata*. Mahwah NJ: Lawrence Erlbaum.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45:5–32.

Bruno, Arpino, Marco Le Moglie, and Letizia Mencarini. 2018. "Machine-Learning techniques for family demography: An application of random forests to the analysis of divorce determinants in Germany." *RECSM Working Paper* 56.

Budig, Michelle J and Paula England. 2001. "The Wage Penalty for Motherhood." *American Sociological Review* 66:204–225.

Cadima, Jorge F. C. L. and Ian T. Jolliffe. 2001. "Variable Selection and the Interpretation of Principal Subspaces." *Journal of Agricultural, Biological, and Environmental Statistics* 6:62–79.

Clark, Andrew, Yannis Georgellis, and Peter Sanfey. 2001. "Scarring: The Psychological Impact of Past Unemployment." *Economica* 68:221–241.

Correll, Shelley J, Stephen Bernard, and In Paik. 2007. "Getting a Job: Is There a Motherhood Penalty?" *American Journal of Sociology* 122:1297–1339.

Daly, Michael J. and Liam Delaney. 2013. "The scarring effect of unemployment throughout adulthood on psychological distress at age 50: Estimates controlling for early adulthood distress and childhood psychological factors." *Social Science and Medicine* 80:19–23.

Dannesfer, Dale. 2003. "Cumulative Advantage Disadvantage and the Life Course: Cross-Fertilizing Age and Social Science Theory." *The Journals of Gerontology: Series B* 58:327–337.

de Linde Leonard, Megan and T. Stanley. 2015. "Married with children: What remains when observable biases are removed from the reported male marriage wage premium." *Labour Economics* 33:72–80.

Degenhardt, Frauke, Stephan Seifert, and Silke Szymczak. 2017. "Evaluation of variable selection methods for random forests and omics data sets." *Briefings in Bioinformatics* 20:492–503.

Doblhammer, Gabriele. 2000. "Reproductive history and mortality later in life: A comparative study of England and Wales and Austria." *Population Studies* 54:169–176.

Elder, Glen H., Monica Kirkpatrick Johnson, and Robert Crosnoe. 2003. "The Emergence and Development of Life Course Theory." In *Handbook of the Life Course*, edited by JeylanT. Mortimer and MichaelJ. Shanahan, Handbooks of Sociology and Social Research, pp. 3–19. Springer US.

Eriksson, Stefan and Dan-Olof Rooth. 2014. "Do Employers Use Unemployment as a Sorting Criterion When Hiring? Evidence from a Field Experiment." *American Economic Review* 104:1014–39.

Falba, Tracy, Hsun-Mei Teng, Jody L Sindelar, and William T Gallo. 2005. "The effect of involuntary job loss on smoking intensity and relapse." *Addiction* 100:1330–1339.

Fasang, Anette Eva. 2012. "Retirement Patterns and Income Inequality." *Social Forces* 90:685–711.

Gabadinho, Alexis, Gilbert Ritschard, Matthias Studer, and Nicolas S. Müller. 2010. "Indice de complexité pour le tri et la comparaison de séquences catégorielles." *Revue des nouvelles technologies de l'information RNTI* E-19:61–66.

Gabriel, Rainer, Michel Oris, Matthias Studer, and Marie Baeriswyl. 2014. "The persistence of social stratification? A life course perspective on old-age poverty in Switzerland." *Swiss Journal of Sociology* (forthcoming).

Gibbons, Robert and Lawrence F Katz. 1991. "Layoffs and lemons." *Journal of Labor Economics* 9:351–380.

Glauber, Rebecca. 2008. "Race and Gender in Families and at Work: The Fatherhood Wage Premium." *Gender & Society* 22:8–30.

Gregg, Paul. 2001. "The Impact of Youth Unemployment on Adult Unemployment in the NCDS." *The Economic Journal* 111:626–653.

Grundy, Emily and Gemma Holt. 2000. "Adult life experiences and health in early old age in Great Britain." *Social Science and Medicine* 51:1061–1074.

Grundy, Emily and Oystein Kravdal. 2007. "Reproductive history and mortality in late middle age among Norwegian men and women." *American Journal of Epidemiology* 167:271–279.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. 2015. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC.

Heckman, James and George Borjas. 1980. "Does Unemployment Cause Future Unemployment? Definitions, Questions and Answers from a Continuous Time Model of Heterogeneity and State Dependence." *Economica* 47:247–83.

Henretta, John C. 2007. "Early childbearing, marital status and women's health and mortality after age 50." *Journal of Health and Social Behavior* 48:254–266.

Henretta, John C. 2010. "Lifetime marital history and mortality after age 50." *Journal of Aging and Health* 22:1198–1212.

Hill, Martha. 1979. "The Wage Effects of Marital Status and Children." *Journal of Human Resources* 14:579–594.

Hofner, Benjamin, Luigi Boccuto, and Markus Göker. 2015. "Controlling false discoveries in high-dimensional situations: boosting with stability selection." *BMC Bioinformatics* 16.

Hofner, Benjamin and Torsten Hothorn. 2017. *stabs: Stability Selection with Error Control*. R package version 0.6-3.

Hoven, Hanno, Nico Dragano, David Blane, and Morten Wahrendorf. 2017. "Early Adversity and Late Life Employment History—A Sequence Analysis Based on SHARE." *Work, Aging and Retirement* 4:238–250.

Hurt, Lisa S, C Ronsmans, and S L Thomas. 2006. "The effect of number of births on women's mortality: Systematic review of the evidence for women who have completed their childbearing." *Population Studies* 60:55–71.

Jolliffe, Ian T. 2002. *Principal Component Analysis*. Springer Series in Statistics., second edition edition.

Kravadal, Oystein, Emily Grundy, T Lyngstad, and K Wiik. 2012. "Family life history and late mid-life mortality in Norway." *Population and Development Review* 38:237–257.

Kursa, Miron B. and Witold R. Rudnicki. 2010. "Feature Selection with the Boruta Package." *Journal of Statistical Software* 36.

Laitinen, Jaana, Ellen Ek, and Ulla Sovio. 2005. "Stress-related eating and drinking behavior and body mass index and predictors of this behavior." *Preventive Medicine* 34:29–39.

Lesnard, Laurent. 2010. "Setting Cost in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns." *Sociological Methods and Research* 38:389–419.

Leung, Man Yee Mallory, Fane Groes, and Raul Santaeulalia-Llopis. 2016. "he Relationship between Age at First Birth and Mother's Lifetime Earnings: Evidence from Danish Data." *PloS one* 11.

Levy, René, Jacques-Antoine Gauthier, and Eric Widmer. 2006. "Entre contraintes institutionnelle et domestique : les parcours de vie masculins et féminins en Suisse." *Cahiers canadiens de sociologie* 31:461–489.

McKetta, Sarah, Seth J. Prins, Jonathan Platt, Lisa M. Bates, and Katherine Keyes. 2018. "Social sequencing to determine patterns in health and work-family trajectories for U.S. women, 1968–2013." *SSM - Population Health* 6:301–308.

McMunn, Anne, Rebecca E Lacey, Meena Kumari, Diana Worts, Peggy McDonough, and Amanda Sacker. 2015. "Work-family life courses and metabolic markers in mid-life: evidence from the British National Child Development Study." *Journal of Epidemiology and Community Health* 70:481–487.

Meinshausen, Nicolai and Peter Bühlmann. 2010. "Stability selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72:417–473.

Miller, Amalia. 2011. "The effects of motherhood timing on career path." *Journal of Population Economics* 24:1071–1100.

OFlaherty, Martin, Janeen Baxter, Michele Haynes, and G Turrell. 2016. "The family life course and health: partnership, fertility histories, and later-life physical health trajectories in Australia." *Demography* 53:1–28.

O'Rand, Angela. 2002. "Cumulative Advantage Theory in Life Course Research." *Annual review of gerontology & geriatrics* 22:14–30.

Orestes Cerdeira, Jorge, Pedro Duarte Silva, Jorge Cadima, and Manuel Minhoto. 2018. *subselect: Selecting Variable Subsets*. R package version 0.14.

Paul, Karsten I and Klaus Moser. 2009. "Unemployment Impairs Mental Health: Meta-Analyses." *Journal of Vocational Behaviour* 74:264–282.

Perry, Chris. 2013. "Machine learning and conflict prediction: a use case." *Stab. Int. J. Secur. Dev.* 2.

Pohlig, Matthias. 2019. "Unemployment sequences and the risk of poverty: from counting duration to contextualizing sequences." *Socio-Economic Review* .

Ribar, David. 2004. "What Do Social Scientists Know About the Benefits of Marriage? A Review of Quantitative Methodologies." *IZA Discussion Paper* 998.

Rossignon, Florence, Matthias Studer, Jacques-Antoine Gauthier, and Jean-Marie Le Goff. 2016. "Childhood family structure and home-leaving: A combination of survival and sequence analyses." In *LaCOSA II*.

Sabbath, Erika L., Ivan Mejía Guevara, M. Maria Glymour, and Lisa F. Berkman. 2015. "Use of Life Course Work–Family Profiles to Predict Mortality Risk Among US Women." *American Journal of Public Health* 105:e96–e102.

Shah, Rajen D. and Richard J. Samworth. 2012. "Variable selection with error control: another look at stability selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75:55–80.

Sironi, Maria. 2018. "Fertility histories and chronic conditions later in life in Europe." *European Journal of Ageing* (forthcoming).

Spivey, Christy. 2005. "Time Off at What Price? The Effects of Career Interruptions on Earnings." *ILR Review* 59:119–140.

Strully, Kate W. 2009. "Job loss and health in the U.S. labor market." *Demography* 46:221–246.

Studer, Matthias. 2013. "WeightedCluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R." LIVES Working Papers 24, NCCR LIVES, Switzerland.

Studer, Matthias. 2018. "Divisive Property-Based and Fuzzy Clustering for Sequence Analysis." In *Sequence Analysis and Related Approaches: Innovative Methods and Applications*, edited by Gilbert Ritschard and Matthias Studer, volume 10 of *Life Course Research and Social Policies*, chapter 13. Springer.

Studer, Matthias, Nicolas S. Müller, Gilbert Ritschard, and Alexis Gabadinho. 2010. "Classer, discriminer et visualiser des séquences d'événements." *Revue des nouvelles technologies de l'information RNTI* E-19:37–48.

Studer, Matthias and Gilbert Ritschard. 2016. "What Matters in Differences between Life Trajectories: A Comparative Review of Sequence Dissimilarity Measures." *Journal of the Royal Statistical Society, Series A* 179:481–511.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 58:267–288.

Tillmann, Robin, Marieke Voorpostel, E Antal, Ursina Kuhn, Florence Lebert, Valérie-Anne Ryser, Oliver Lipps, and Wernli Boris. 2016. "The Swiss Household Panel Study: Observing social change since 1999." *Longitudinal and Life Course Studies* 7:64–78.

Vanhoutte, Bram, Morten Wahrendorf, and Jacques Yzet Nazroo. 2017. "Duration, timing and order: How housing histories relate to later life wellbeing." *Longitudinal and Life course Studies* 8.

Widmer, Eric D., René Levy, Alexandre Pollien, Raphaël Hammer, and Jacques-Antoine Gauthier. 2003. "Between Standardisation, Individualisation and Gendering: An Analysis of Personal Life Courses in Switzerland." *Swiss Journal of Sociology* 29:35–65.

Williams, Krist, Sharon Sassler, Arianne Frech, Fenaba Addo, and Elizabeth Cooksey. 2011. "Nonmarital childbearing, union history, and women's health at midlife." *American Sociological Review* 76:465–486.

Yuan, Ming and Yi Lin. 2006. "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68:49–67.

Zaki, Mohammed Javeed. 2001. "SPADE: An Efficient Algorithm for Mining Frequent Sequences." *Machine Learning* 42:31–60.

Zoutewelle-Terovan, Mioara, Victor van der Geest, Aart Liefbroer, and Catrien Bijleveld. 2012. "Criminality and Family Formation." *Crime & Delinquency* 60:1209–1234.

# A Boruta results for men

Table 8: Features selected by Boruta for income and self-rated health for men.

Family Trajectories.

| Feature | Income | | Health | |
|---|---|---|---|---|
| | Imp. | Prop. | Imp. | Prop. |
| **Sequencing** | | | | |
| Oth → Part/Kid | | | 3.82 | 0.70 |
| Part/Kid | 2.83 | 0.65 | 4.62 | 0.86 |
| Part | | | 3.99 | 0.75 |
| Part → Part | 3.48 | 0.80 | | |
| Part → Sin | 3.09 | 0.73 | 5.19 | 0.91 |
| Part → Sin → Part | | | 3.72 | 0.67 |
| Sin → Oth | | | 3.79 | 0.70 |
| Sin → Oth → Part/Kid | | | 4.02 | 0.75 |
| Sin → Part/Kid | | | 4.03 | 0.75 |
| Sin → Part | | | 4.20 | 0.81 |
| Sin → Part → Part | 3.87 | 0.88 | | |
| Sin → Part → Sin | 3.58 | 0.81 | 5.69 | 0.94 |
| Sin → Sin | | | 4.63 | 0.87 |
| **Timing** | | | | |
| 20-24 Oth | | | 4.26 | 0.78 |
| 20-24 Part | 3.64 | 0.83 | | |
| 20-24 Sin | 2.98 | 0.67 | 5.11 | 0.89 |
| 25-29 Oth | | | 5.43 | 0.92 |
| 25-29 Part/Kid | 5.30 | 0.98 | 7.61 | 1.00 |
| 25-29 Part | 3.10 | 0.72 | 4.82 | 0.85 |
| 25-29 Sin | | | 5.05 | 0.90 |
| 30-34 Oth | | | 5.82 | 0.95 |
| 30-34 Part/Kid | 3.11 | 0.67 | 7.25 | 1.00 |
| 30-34 Part | 3.66 | 0.85 | 6.35 | 0.98 |
| 30-34 Sin | | | 4.20 | 0.77 |
| **Duration** | | | | |
| Overall Oth | | | 7.41 | 0.98 |
| Overall Part/Kid | 5.54 | 0.98 | 10.30 | 1.00 |
| Overall Part | 5.21 | 0.98 | 8.94 | 1.00 |
| Overall Sin | 3.82 | 0.86 | 8.81 | 1.00 |
| **Complexity** | | | | |
| Complexity Index | 4.02 | 0.90 | 9.63 | 1.00 |

Work Trajectories.

| Feature | Income | | Health | |
|---|---|---|---|---|
| | Imp. | Prop. | Imp. | Prop. |
| **Sequencing** | | | | |
| Edu → FT → PT | 3.71 | 0.83 | | |
| FT → NW | | | 5.85 | 0.95 |
| NW | | | 5.45 | 0.92 |
| NW → FT | | | 3.44 | 0.59 |
| **Timing** | | | | |
| 20-24 Edu | 4.76 | 0.95 | | |
| 25-29 Edu | 3.96 | 0.86 | | |
| 30-34 FT | 4.04 | 0.88 | | |
| 30-34 NW | | | 7.12 | 0.99 |
| **Duration** | | | | |
| Overall Edu | 5.24 | 0.97 | | |
| Overall FT | 3.24 | 0.72 | | |
| Overall NW | | | 6.21 | 0.97 |
| Overall PT | 3.69 | 0.85 | | |

# B Comparison Boruta and Feature Selection results for men

Table 9: Features selected by Boruta for income and self-rated health for men.

| Income | |
|---|---|
| Feature | |
| **Sequencing** | |
| *Work Trajectories* | |
| Edu → FT → PT | |
| *Family Trajectories* | |
| Part/Kid | S |
| Part → Part | S |
| Part → Sin | |
| Sin → Part → Part | |
| Sin → Part → Sin | |
| **Timing** | |
| *Work Trajectories* | |
| 20-24 Edu | A |
| 25-29 Edu | A |
| 30-34 FT | * |
| *Family Trajectories* | |
| 20-24 Part | |
| 20-24 Sin | |
| 25-29 Part/Kid | |
| 25-29 Part | |
| 30-34 Part/Kid | |
| 30-34 Part | |
| **Duration** | |
| *Work Trajectories* | |
| Overall Edu | * |
| Overall FT | |
| Overall PT | |
| *Family Trajectories* | |
| Overall Part/Kid | A |
| Overall Part | S |
| Overall Sin | |
| **Complexity** | |
| *Work Trajectories* | |
| Complexity Index | |

| Health | |
|---|---|
| Feature | |
| **Sequencing** | |
| *Work Trajectories* | |
| FT → NW | S |
| NW | S |
| NW → FT | A |
| *Family Trajectories* | |
| Oth → Part/Kid | |
| Part/Kid | |
| Part | A |
| Part → Sin | |
| Part → Sin → Part | |
| Sin → Oth | |
| Sin → Oth → Part/Kid | |
| Sin → Part/Kid | |
| Sin → Part | |
| Sin → Part → Sin | |
| Sin → Sin | |
| **Timing** | |
| *Work Trajectories* | |
| 30-34 NW | * |
| *Family Trajectories* | |
| 20-24 Oth | |
| 20-24 Sin | |
| 25-29 Oth | |
| 25-29 Part/Kid | S |
| 25-29 Part | S |
| 25-29 Sin | |
| 30-34 Oth | A |
| 30-34 Part/Kid | |
| 30-34 Part | S |
| 30-34 Sin | |
| **Duration** | |
| *Work Trajectories* | |
| Overall NW | A |
| *Family Trajectories* | |
| Overall Oth | |
| Overall Part/Kid | |
| Overall Part | S |
| Overall Sin | |
| **Complexity** | |
| *Family Trajectories* | |
| Complexity Index | |

Note: For the abbreviation of the state, please refer to Table 1.
* = features selected by both methods; A = direct linear relationship is accounted for when controlling; S = direct linear relationship remains significant after controlling; empty = direct linear relationship is not significant in any case.

# C Regression results for men

Table 10: Logistic regressions on health for men.

| | Base | Seq. An. | Stab. Sel. | Complete |
|---|---|---|---|---|
| Intercept | $2.85^{***}$ | $2.91^{***}$ | $3.21^{***}$ | $3.14^{***}$ |
| | (0.66) | (0.70) | (0.68) | (0.71) |
| Age | $-0.03^{*}$ | $-0.03^{*}$ | $-0.03^{**}$ | $-0.03^{*}$ |
| | (0.01) | (0.01) | (0.01) | (0.01) |
| Tertiary educ. | $0.47^{**}$ | $0.50^{**}$ | $0.47^{**}$ | $0.48^{**}$ |
| | (0.18) | (0.18) | (0.18) | (0.18) |
| Born in CH | $0.49^{**}$ | $0.47^{**}$ | $0.46^{**}$ | $0.44^{*}$ |
| | (0.17) | (0.18) | (0.18) | (0.18) |
| 30-34 NW | | | $-0.71^{***}$ | $-0.79^{**}$ |
| | | | (0.18) | (0.28) |
| Professional Cluster (ref=Full-time) | | | | |
| Late FT-NW | | $-1.93^{**}$ | | $-0.38$ |
| | | (0.74) | | (0.98) |
| FT-NW | | $-2.19$ | | 1.72 |
| | | (1.48) | | (2.01) |
| PT | | $-0.18$ | | $-0.15$ |
| | | (0.40) | | (0.40) |
| Late FT-PT | | $-0.26$ | | $-0.27$ |
| | | (0.56) | | (0.56) |
| Edu-FT | | $-0.03$ | | $-0.05$ |
| | | (0.26) | | (0.26) |
| NA | | $-0.30$ | | $-0.34$ |
| | | (0.79) | | (0.79) |
| Cohabitation Cluster (ref=Single) | | | | |
| Early parenthood | | $-0.20$ | | $-0.20$ |
| | | (0.26) | | (0.27) |
| With partner | | 0.40 | | 0.41 |
| | | (0.33) | | (0.34) |
| Other paths | | $-0.42$ | | $-0.47$ |
| | | (0.35) | | (0.35) |
| Late parenthood | | 0.06 | | 0.06 |
| | | (0.24) | | (0.24) |
| AIC | 910.76 | 916.34 | 897.23 | 909.75 |
| BIC | 931.06 | 987.42 | 922.61 | 985.90 |
| Log Likelihood | $-451.38$ | $-444.17$ | $-443.62$ | $-439.88$ |
| $\Delta\chi^2$ (vs Base) | | 14.41 | $15.53^{***}$ | $23.00^{*}$ |
| $\Delta\chi^2$ (vs Complete) | $23.00^{*}$ | $8.59^{**}$ | 7.48 | |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 11: Linear regression on standardized household income for men.

| | Base | Seq. An. | Stab. Sel. | Complete |
|---|---|---|---|---|
| Intercept | 23828.06* | 23978.86* | 15736.06 | 19533.85 |
| | (10141.48) | (10415.81) | (10743.23) | (13866.84) |
| Age | 617.02*** | 652.04*** | 550.45** | 584.60** |
| | (185.16) | (184.70) | (182.77) | (185.26) |
| Tertiary educ. | 21089.93*** | 17798.61*** | 17133.64*** | 17115.24*** |
| | (2815.74) | (2827.91) | (2811.20) | (2823.09) |
| Born in CH | 5931.39* | 7695.88** | 8718.30** | 8369.80** |
| | (2729.46) | (2722.93) | (2732.32) | (2754.07) |
| Overall Edu | | | 3593.51*** | 2973.84* |
| | | | (1024.47) | (1311.61) |
| 30-34 FT | | | 1796.30 | 1427.80 |
| | | | (979.36) | (1882.32) |
| NW | | | −15769.75* | −14866.84* |
| | | | (6300.30) | (6879.35) |
| Edu → FT | | | 5144.14 | 3467.91 |
| | | | (5113.73) | (5255.09) |
| Professional Cluster (ref=Full-time) | | | | |
| Late FT-NW | | −9976.19 | | 8183.00 |
| | | (17784.21) | | (19254.94) |
| FT-NW | | −45351.73 | | −31127.89 |
| | | (27235.83) | | (29160.91) |
| PT | | −5591.29 | | −1827.54 |
| | | (5842.43) | | (10307.74) |
| Late FT-PT | | −1364.23 | | 1731.20 |
| | | (8728.32) | | (10225.14) |
| Edu-FT | | 23910.12*** | | 6176.80 |
| | | (3766.40) | | (6981.74) |
| NA | | −11621.80 | | −5341.30 |
| | | (13117.06) | | (15004.19) |
| Cohabitation Cluster (ref=Single) | | | | |
| Early parenthood | | −6541.92 | | −6148.36 |
| | | (4153.91) | | (4137.29) |
| With partner | | −244.63 | | 484.98 |
| | | (4529.64) | | (4531.78) |
| Other paths | | −6388.10 | | −5814.24 |
| | | (5901.85) | | (5890.23) |
| Late parenthood | | −5995.59 | | −5088.41 |
| | | (3606.73) | | (3604.00) |
| $R^2$ | 0.05 | 0.09 | 0.10 | 0.10 |
| $\Delta F$ (vs Base) | | 5.50*** | 15.96*** | 5.05*** |
| $\Delta F$ (vs Complete) | 5.05*** | 3.80*** | 0.70 | |
| BIC | 30996.03 | 31013.07 | 30961.95 | 31026.34 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$