Please cite as: Ritschard, G, R. Bürgin and M. Studer (2013). Exploratory Mining of Life Event Histories. In J.J. McArdle & G. Ritschard (eds), *Contemporary Issues in Exploratory Data Mining in Behavioral Sciences*, Routeledge, New York, pages 221–253

Exploratory Mining of Life Event Histories

Gilbert Ritschard, Reto Bürgin and Matthias Studer

Abstract

This chapter explains how data-mining-based techniques can be used for discovering interesting knowledge from sequences of life events, that is, to find out how people sequence important life events. We illustrate with data from the biographical survey conducted by the Swiss Household Panel in 2002. The focus is on the sequencing of events in the occupational life course and of events such as starting a union and childbirth that affect the living arrangement. Addressed methods include finding of frequent sequential patterns, identification of discriminant subsequences and clustering of event sequences.

1 Introduction

Mining of sequentially organized data has been successfully exploited in many domains such as genetics, device control, speech recognition as well as for automatic text, customer behavior and web logs analysis. In the social sciences sequence exploration mainly focuses on state sequences, and consists typically in building typologies by means of the optimal matching approach (Abbott and Forrest, 1986). In contrast, we focus in this chapter on event sequences rather than state sequences. Let us clarify the difference between the two.

A state, such as being jobless, lasts the whole considered unit of time while an event, for example ending a job, occurs at a certain time point and has no duration. The event does not last, but provokes, possibly in conjunction with other events, a state change. In state sequences, the positions in the sequence reflect the duration since the beginning of the sequence, while in event sequences they just inform about the number of precedent events. Therefore, while state sequences are particularly of interest for studying durations and timing in life courses, event sequences are especially useful when the concern is the sequencing, i.e., the order in which events occur. This does not mean that we cannot account for the time of occurrence of the events. However, since position does not convey time information, explicit time stamps are needed for that. Another important difference between states and events is that multiple events can occur at a same time point, while states are mutually exclusive. We cannot be in two different states at the same time. Further-

more, depending on the situation at the time the event occurs, a same event may characterize different transitions. For example, when living alone the event 'having a childbirth' provokes the transition from state 'Alone' to state 'Alone with a child' while for someone married it provokes the transition from state 'Married' to the state 'Married with a child'. Although there is not always a clear univocal relationship between state sequences and event sequences, each of them is just an alternative way of looking at the same information about life trajectories. Event sequences need different tools than state sequences. For example, while state sequences are easily rendered with stacked segments of colored lines each representing the time spent in a given state, events cannot be rendered this way because they have no duration.

So, our aim in this chapter is to show what we can do with event sequences and the kind of results that we can expect from an event sequence analysis. While state sequences have received a lot of attention in the social sciences, especially since the popularization of the optimal-matching-based methods by Abbott in the late 80's (Abbott and Forrest, 1986; Aisenbrey and Fasang, 2010), and there exist nowadays efficient pieces of software to explore such state sequences (e.g., Brzinsky-Fay et al., 2006; Gabadinho et al., 2011), event sequence analysis has received much less attention in the social science.

The approach followed in the few social science papers that consider event sequences (Abbott, 1983, 1991; Blockeel et al., 2001; Billari et al., 2006) mainly consists, with the noticeable exception of Blockeel et al. (2001), in looking at the frequencies of a priori defined subsequences of interest. Here,

we adopt a more holistic point of view and explore all subsequences that can be found in the data. We essentially place our approach in the line of the work developed by the data mining community (Agrawal and Srikant, 1995; Mannila et al., 1995; Bettini et al., 1996; Mannila et al., 1997; Zaki, 2001) as an extension of the mining of frequent—non ordered—patterns. Behind these aspects, we also consider the measure of pairwise dissimilarities between event sequences, which then gives access to any dissimilarity-based method.

Broadly, the addressed methods consist in finding the most frequent subsequences, i.e., the most common ways of sequencing life events, and then in finding out among them those that best discriminate between groups such as women and men for example. The measure of pairwise dissimilarities between event sequences is also addressed and we show how such dissimilarities can be used for clustering event sequences. Regarding the event time-stamps, we show that, when available, they can be used for restricting the search for interesting subsequences through time constraints. Differences in the timing can also explicitly be accounted for when computing dissimilarities between time-stamped event sequences.

It is worth mentioning that the approach proposed here differs from traditional event history analysis (Blossfeld et al., 2007; Yamaguchi, 1991), which is indeed survival analysis and as such focuses on the timing or hazard of only one specific event, the death, the marriage or the first job for example. See Ghisletta (2012), Scott et al. (2012) and Zhou et al. (2012) in this volume for applications using exploratory data mining approaches in a survival analysis

context.

For the social sciences, the exploration of event sequences should permit to answer questions such as

- What is the most typical succession of family or professional life events?
- Are there standard ways of sequencing those events?
- What are the most typical events that occur after a given subsequence such as after leaving home and ending education?
- How is the sequencing of events related to covariates?
- Which event sequencings do best discriminate groups such as men and women?

As already mentioned, our aim is to demonstrate the scope of sequential event pattern mining in social sciences. We exploit for that the features offered by the TraMineR package (Gabadinho et al., 2011) for event sequences[1] and illustrate with data on Swiss cohabitational and occupational life courses. TraMineR is a general toolbox for exploring state, as well as, event sequence data in the free open source graphical and statistical environment R (R Development Core Team, 2012). The name TraMineR stands for 'Trajectory Miner in R'.

[1]See Studer et al. (2010) and the user's guide (Gabadinho et al., 2009) for details about TraMineR 's event sequence features.

2 The data

We consider sequences derived from the biographical survey conducted in 2002 by the Swiss Household Panel.[2] We retain the 1503 cases studied in Widmer and Ritschard (2009) with techniques for state sequences, but look at the events behind the state changes. The retained individuals were all aged 45 years or more at the survey time, i.e., they were born in 1957 or earlier, and the focus is on their life trajectories between 20 and 45 years, the data granularity being at the yearly level. Table 1 gives the alphabet of the cohabitational and occupational states analyzed in Widmer and Ritschard (2009) and Figure 1 shows the index-plot—rendering of the individual sequences—of each of the two considered sets of state sequences. In those plots, each sequence is represented by a horizontal line colored according to the state at the successive positions. Color changes indicate state transitions and the length of each segment in a same color reflects the time spent in the corresponding state. In Figure 1 the sequences are sorted by states from the end to the beginning of the time frame.

[Table 1 about here.]

The data were selected on the basis of fully informed cohabitational trajectories but incomplete sequences of occupational sequences were not excluded. There are even some cases for which we have no information at all about occupation. These most probably correspond to people who never

[2] `www.swisspanel.ch`

worked and stayed at home during the whole observed age interval, i.e., from 20 to 45 years old.

[Figure 1 about here.]

[Table 2 about here.]

Cohabitational event sequences are derived from the cohabitational state sequences by specifying the events that cause the transition between any two states. The list of considered cohabitational events is given in Table 2 and the retained definition of the transitions in terms of those events is depicted in Table 3. To give an example of how this latter table should be read, consider the list `"LH,A"` at the intersection of row '2P' and column 'A'. This list of two events indicates that each time we observe a transition from the state '2P' (leaving with both parents) to the state 'A' (living alone) we assume that the concerned individual experienced the events `LH` ('leaving home') and `A` ('start living alone'). The terms on the diagonal of the table have a somewhat different meaning. They indicate the event that we assign when the state sequence starts in the corresponding row state. Those diagonal terms serve thus only at the beginning of the sequences. For example, the `"2P"` in the first cell of the table means that we assign the event `"2P"` at age 20 for all sequences starting in state 'living with both parents' (2P).

For occupational trajectories, we define the events as the start of the spells in a same state. We do so, however, after having assimilated the states 'missing' (Mi), for the reason mentioned above, and 'retired' (RE), which

is observed only four times, to the state 'at home' (AH). We thus get six occupational events, namely starting one of the following: an at home spell, working full time, working part time, a negative break, a positive break and an education period, which we respectively denote as AH, FT, PT, NB, PB and ED.

[Table 3 about here.]

[Figure 2 about here.]

The categorical parallel coordinate plots (Bürgin et al., 2012) in Figure 2 show the diversity of event sequencing in the considered cohabitational and occupational trajectories. The diversity is rendered by avoiding overlapping thanks to small displacements of the lines. The figure also highlights with colors the most typical patterns, that is, here, those that are shared by at least 5% of the observed cases. In that graphic, sequences which form a prefix of another one, i.e., which are identical with the beginning of some other longer sequence, are merged into this longer sequence, and the plot renders the frequencies of events and embedded sequences with varying width of both the event points and the connecting segments between successive events. Let us look at the the lower graphic of Figure 2 to clarify how the plots should be read. We first notice that the successive event squares of a given pattern are always located at the same position in the light-grey zones drawn as background of the coordinate points. This facilitates tracking the successive events of a same sequence. Now, let us consider the green pattern. Since

the green square at position 1 is clearly larger than the one at position 2, it follows that a large proportion of the observed people experienced only the first event, i.e., 'start working full time' (FT). Likewise, since the second green segment is thinner than the first one, it exhibits that only a fraction of those who start staying at home (AH) after having started to work full time, return later working part time (PT).

In the upper plot in Figure 2, we observe, for instance, that most of the considered individuals were living with their both parents (2P) at the beginning of the sequence, i.e., when they were 20 years old and that leaving home (LH) and moving in together with a partner (U) in the same year is very common in the 20th century Switzerland. It is also common to have the first childbirth in this same year. More spaced events are quite frequent too, but the plot clearly shows that the standard order of the events is leaving home, starting a union and then having the first childbirth.

Regarding occupational trajectories, we see that the most frequent start points at 20 years old are education (ED) and working full time (FT) and that it is very common to switch directly from education to full-time work and from full-time work to an at home stay (AH). We also observe that after switching from working full-time to an at home stay, it is quite current to return working part-time (PT).

3 Frequent subsequences vs frequent itemsets

The mining of frequent itemsets and association rules has been popularized in the 90's with the work of Agrawal and Srikant (1994) and Agrawal et al. (1995) and their *Apriori* algorithm. The work was for instance intended to analyze "buying baskets," i.e., to find out items that customers often buy together and rules such as "if a buyer buys a product A then he or she will most probably also buy B", or to the control of devices where the interest is to discover symptoms that often occur together and announce a failure. Adopting a longitudinal perspective, the methods were then extended to account for the sequences of purchases or of symptom occurrences, and this is where the mining of frequent subsequences started.

The main idea of *Apriori* is to exploit the property that an itemset of size $k+1$ obtained by adding an item to a non-frequent itemset of size k cannot be frequent. It considers thus successively only itemsets of size $k+1$ derived from frequent itemsets of size k, which reduces considerably the number of itemsets to scan.

Mining typical event sequences is in some sense a specialized case of the mining of frequent itemsets. It is much more complex, however, and requires the user to specify time constraints and select a counting method. While there is a general agreement about how to count occurrences of itemsets in the classical unordered framework, there is no such agreement for episodes

(subsequences) that may occur more than once in the same sequence. The additional time dimension raises questions such as: What is the maximum time span, i.e., sequence length we want to analyze? Until which time gap should events be considered to occur together? Joshi et al. (2001) nicely present the various counting schemes and possibly useful time constraints. A discussion of those aspects in a social science perspective can be found in Ritschard et al. (2008) and we shortly recall some of the main points in the next section.

Efficient algorithms for extracting frequent subsequences have been proposed in the literature among which the prominent ones are those of Bettini et al. (1996), Srikant and Agrawal (1996), Mannila et al. (1997) and Zaki (2001). The algorithm implemented in TraMineR is an adaptation of the prefix-tree-based search described in Masseglia (2002).

4 Mining frequent sequential patterns

We show in this section how we can search for frequent subsequences in cohabitational and occupational event trajectories and the kind of knowledge we can gain from them. We also provide examples of the basic TraMineR commands used for extracting the frequent patterns.

Before turning to the frequent subsequences, it is worth to first specify the terminology— from Studer et al. (2010)—which we use, since there is no clear standard about it in the literature.

4.1 Terminology

Formally, as in Studer et al. (2010), we define a *transition* as a—non ordered—set of events occurring at the same time point, namely the set of events that cause the state transition. For instance the joint occurrence of the two events 'leaving home' and 'moving in with a partner' results in the transition from the state 'living with both parents' (2P) or possibly from 'living with one parent' (1P) to the state 'living with a partner' (U). A transition is also known in the data mining literature as a transaction, especially when the concern is the mining of customer behaviors.

An *event sequence* is then defined as an ordered list of transitions. We represent it as a succession of transitions separated by edges or arrows. Thus, in the sequence

$$\texttt{(LHome, Union)} \rightarrow \texttt{(Marriage)} \rightarrow \texttt{(Childbirth)} \tag{1}$$

the first term `(LHome, Union)` is a transition defined by the two events `LHome` and `Union`, while the two next transitions result each from a single event. To account for the time gap between two consecutive transitions, we add *duration stamps* when necessary. With the following notations

$$\xrightarrow{22} \texttt{(LHome, Union)} \xrightarrow{3} \texttt{(Marriage)} \xrightarrow{2} \texttt{(Childbirth)} \xrightarrow{15}$$

we indicate that the considered individual gets married 3 years after moving

in together with the partner and has a childbirth 2 years after marriage. We also indicate that he or she leaves home to move in with a partner 22 years after the start of observation (birth), and that he or she remains under observation for 15 years after the childbirth without experiencing any other event of interest.

An *event subsequence* y of an event sequence x, is an event sequence such that all its events also belong to x and occur in the same order as in x. For instance, `(Lhome)`→`(Childbirth)` is a subsequence of the example sequence (1) while `(Lhome)`→`(Union)` is not a subsequence of it.

The notion of state subsequence, as used for instance by Elzinga (2003, 2010), would more or less correspond to a subsequence of transitions. Event subsequences are different in that each transition may contain more than one event and that a same event may be present in different transitions. For our data sets, event and state subsequences would be equivalent for the occupational trajectories for which the events indicate the start of each spell in a same state. They are, however, quite different for the cohabitational trajectories where we used the more complex transformation defined by Table 3.

4.2 Inputting data to TraMineR

Data can be provided to TraMineR in the vertical time-stamped event form, i.e., with a different line for each event. The data would then look as shown in Table 4.

[Table 4 about here.]

In that example, the time stamp is the age at which the individual experiences the event and we can see that individual 101 both leaves parental home and moves in together with a partner at the age of 22.

We may also provide state sequences and select one of several methods for converting them into event sequences (see Gabadinho et al., 2009; Ritschard et al., 2009, for details). We adopt this solution for our illustrative data since they are organized as state sequences. For the cohabitational trajectories, we select the method consisting in associating to each state transition the events as specified by the transition definition Table 3. Letting for instance `seqs.coh` be the state sequence object for the cohabitational trajectories and `transition.coh.mat` the transition definition matrix for cohabitation trajectories given in Table 3, we derive the event sequence object from them with the following simple command

```
R> shpevt.coh <- seqecreate(seqs.coh, tevent=transition.coh.mat)
```

Table 5 shows how the first 5 event sequences look. Notice that each sequence has a start event corresponding to the element given on the corresponding diagonal element of the transition definition table.

[Table 5 about here.]

For the occupational trajectories, we first recoded states Mi and RE as AH and then generated the event sequences with the option `tevent="state"`,

which assigns an event to the start of each spell in a same state and names the event with the name of that state.

```
R> shpevt.occ <- seqecreate(seqs.occ, tevent="state")
```

4.3 Finding the frequent subsequences

As already mentioned above, a given subsequence may occur more than once in the same sequence and we have to chose a counting method to determine its frequency. In TraMineR we can chose from Joshi et al. (2001)'s six different counting methods. The first and perhaps most common one (COBJ) is to count the number of sequences that contain the subsequence of interest. The second one (CWIN) sets a sliding window size and counts in each sequence the number of windows that contain the subsequence and then adds up the counts. The third one (CMINWIN) proceeds similarly but with using in each sequence the smallest window size that contains the subsequence. The last two methods consist in counting the number of occurrences of the subsequence in each sequence, allowing for possible overlaps of the found occurrences in the fourth method (CDIST_O) and considering only non-overlapping occurrences in the fifth one (CDIST). In the last case, the solution may depend on the order in which we constitute the successively counted subsequences.

4.3.1 Most frequent cohabitational subsequences

We now extract the cohabitational subsequences supported by at least 50 cases, which we obtain with the COBJ counting method. The TraMineR command is

```
R> cons <- seqeconstraint(countMethod='COBJ')
R> shp.fss.coh <- seqefsub(shpevt.coh, minSupport=50,
+       constraint=cons)
```

We get 85 subsequences. The most frequent ones contain a single event, which is not very instructive about event sequencing. Therefore, we display in Table 6 the 10 most frequent subsequences that contain at least two events.

[Table 6 about here.]

[Table 7 about here.]

We observe that the most frequent subsequences consist of only four out of the ten considered events, namely: living with both parents, leaving home, starting a union and having a first childbirth. We learn from Table 6 that the most frequent and hence most typical event sequencing is to live with both parents and then leave home more than a year later. This succession of events is experienced by 62.1% of the 1503 cases, while, after living with both parents when being 20 years old or more, 58.2% move in with a partner and 47.7% have a first childbirth. People who follow the 6th most frequent subsequence (2P) $\rightarrow$ (LH,U) are also counted among those who experienced

the most frequent two subsequences (2P) $\rightarrow$ (LH) and (2P) $\rightarrow$ (U). We thus deduce that there are $345 = 934 - 589$ individuals who left home without moving in with a partner at the same time. Other similar results can be derived from Table 6. For instance, if we look at the 7th and 8th most frequent subsequences, we can establish that 0.6% of the individuals in the sample started a union, had a first child more than a year afterwards, but did not live at any time with both parents after they were 20 years old.

We get very similar results by counting distinct occurrences (CDIST_O) of the subsequences rather than the number of sequences that include them. It leads to an increase of roughly 3 to 5% of the count support. This is not surprising since most of the considered events occur only once in each sequence. It is more interesting to vary the maximal time span allowed for the subsequences. Setting the maximum time span as three years, we get only 29 subsequences with a support of 50 or more. The 10 most frequent are displayed in Table 7 and comparing with Table 6 we can see some changes in the order of the subsequences. Leaving home to start a union, (LH,U), and having the first childbirth when moving in with a partner, (C,U), are the most frequent pattern occurring within three years. There are also important drops in support. For example, only 412 cases out the 645 who experienced the sequence (U) $\rightarrow$ (C), had their first childbirth within 3 years after they started to live with a partner.

4.3.2 Most frequent occupational subsequences

The 10 most frequent occupational subsequences with more than one event are listed in Table 8. The sequencing of occupational trajectories clearly looks less standardized since the frequencies of the most frequent subsequences are about three times lower than those of the most frequent cohabitational trajectories. Remember that the events here are just the transitions between states and that we have assimilated the start of a retired (RE) or missing (Mi) spell to the start of an at home (AH) spell. There are no simultaneous events.

[Table 8 about here.]

We learn from Table 8 that three transitions are experienced by more than 20% of the cases, namely starting eduction and later a full-time job (28.3%), working full-time and then staying at home (26.5%) and starting a full-time job and later working part time (21.9%). Among those who start an at home stay after having worked full time, 60.3% (the 158 cases who experienced the 8th most frequent subsequence) return working part-time afterwards. These are, as we will see below, mainly women.

[Table 9 about here.]

As for cohabitational trajectories, we get more or less the same most frequent subsequences with all counting methods. With CDIST_O, the counting supports increase by about 6 to 10%, i.e., a bit more than for cohabitational

subsequences. This indicates that repeating events are slightly more frequent in occupational trajectories than in cohabitational ones. Let us look at subsequences with a three-year maximum time span. Table 9 reports the only six such subsequences that satisfy the minimum count-support of 50. Comparing with Table 8 we see that while 68% of those who started to work full time after education, (ED) $\rightarrow$ (FT), did it within 3 years from the beginning of the education spell—which is most often also the beginning of the sequence—, only 25% of those who stopped working at full time to stay at home, (FT) $\rightarrow$ (AH), did it within 3 years from starting to work full time.

4.3.3 Merging cohabitational and occupational channels

It is also interesting to combine both sets of events and to consider joint event sequences. Table 10 shows the most frequent subsequences of the combined cohabitational and occupational trajectories.

[Table 10 about here.]

Unsurprisingly, in this list we find the most frequent cohabitational subsequences that are more than twice as frequent as the most frequent occupational subsequences. However, there are also some subsequences that combine cohabitational and occupational events. Among them, the two most frequent are moving in with a partner after having started a full-time work and having the first childbirth after having started a full time work. They are respectively shared by 69.5% and 58.3% of the cases. We also observe that

it is common to leave home after having started to work full time (55.5%), and even to leave home and move in with a partner in the same year after starting to work full time (37.6%).

With the previous exploration, we have been able to identify the standard ways of sequencing the life events of interest. The next step would be to study whether there are differences in those standards among socio-demographic groups such as those defined by gender or birth cohort. A possible solution would be to look for the most frequent subsequences separately in each group and then compare them. It is more efficient, however, to directly seek the most discriminant subsequences.

5 Discriminant subsequences

When examining the most typical sequential patterns, questions naturally arise about their relationship with covariates such as sex or birth cohort. For instance, we may be interested to know which sequence pattern best characterizes women or youngest cohorts.

To answer those questions, we use the method proposed in Studer et al. (2010) which consists of measuring the strength of association of each subsequence with the considered covariate and then selecting the subsequences with the strongest association. The association can be measured with the Pearson independence Chi-square. We define for this a 0-1 presence indicator variable of the subsequence, cross tabulate it with the covariate and

compute the Pearson Chi-square of the resulting table. The tables are of the same size for all subsequences since they all cross tabulate a binary indicator variable with a same covariate. Hence, the Chi-square can be directly used for sorting the subsequences, the most discriminant one being the one with the highest Chi-square. We would get the same order with a normalized Chi-square such as Cramer's v or by sorting in increasing order of the associated p-values for the independence test.[3]

5.1 Differentiating between sexes

Below is the TraMineR command with which we obtain, in decreasing order of their discriminant power, the cohabitational subsequences that best distinguish between sexes. We limit the search to subsequences which discriminate sex at the 1% significance level.

```
R> shp.dss.coh <- seqecmpgroup(shp.fss.coh, group = seqs$sex,
+       pvalue.limit = 0.01)
```

The 6 most discriminating cohabitational subsequences are listed in Table 11 and the frequencies of all 13 subsequences that significantly discriminate for the sex at the 1% level are plotted in Figure 3.

[Table 11 about here.]

[3]Although, as pointed out to us by Raffaella Piccarreta, directional measures such as the τ of Goodman and Kruskal (1954) or the uncertainty coefficient u of Theil (1970) would be better suited for this discrimination purpose, they are not yet implemented in TraMineR.

[Figure 3 about here.]

The colors used for the bars in Figure 3 indicate the sign and significance of the associated Pearson residual. This residual is the signed square root of the contribution to the Chi-square of the cell $(1, g)$ corresponding to the presence, 1, of the subsequence in the group g

$$\text{Pearson residual } = \frac{(n_{1g} - e_{1g})}{\sqrt{e_{1g}}}$$

where n_{1g} is the observed count and e_{1g} the expected count under the independence assumption. At first glance, it may be surprising to get non significant—small—Pearson residuals for discriminant subsequences. As can be seen in Figure 3, this happens for highly frequent subsequences such as those formed by the single event 'starting to live with both parents' (2P) or 'moving in with a partner' (U). The reason is that in such cases, the negation of a frequent subsequence, e.g., 'never living with both parents' or 'never moving in with a partner', will be small and the associated Pearson residual very large.

Although there are significant differences between men and women in the frequencies of some cohabitational subsequences, the differences are not very important and are probably due to timing effects that are hidden because we do not account for events occurring before the age of 20. For instance, the lower proportion of women that leave home between ages 20 and 45 can be explained by the higher proportion of women who leave home before their

20th birthday.

[Table 12 about here.]

[Figure 4 about here.]

Differences between men and women are much more important in the sequencing of occupational events. Table 12 lists the 6 most discriminant occupational subsequences while Figure 4 plots the frequencies of all subsequences which are significantly discriminant at the 0.1% level.

Interestingly, most of the occupational discriminant subsequences are quite frequent for women and rarely observed for men. The three exceptions are $\mathrm{ED}, \mathrm{ED} \rightarrow \mathrm{FT}$ and $\mathrm{FT} \rightarrow \mathrm{ED} \rightarrow \mathrm{FT}$, which indicate that more men than women start an education spell after 20 years old, that men more often start working full time after education, and that men more often return to working full time after an education break. All the other discriminant subsequences contain either the 'at home', AH, or the 'partial time', PT, event, which, in 20th century Switzerland, typically are events experienced by women. The results clearly demonstrate this fact.

As we did it with frequent subsequences, it is interesting to look for discriminant subsequences of trajectories that combine the cohabitational and occupational events of each individual. The 10 most discriminant subsequences of the mixed sequences are listed in Table 13 and we can see that all of them contain one of the two events AH and PT which are typically experienced by women. In four cases, at least one of these two events is combined

with 'moving in with a partner' (U) and in one case with 'first childbirth' (C). The results clearly demonstrate that staying at home or switching to part-time work sometime after moving in with a partner is a typically female behavior. This is also true for switching to part-time work sometime after the first childbirth.

[Table 13 about here.]

5.2 Differentiating among birth cohorts

Let us now look at differences in the event sequencing among birth cohorts. We consider three cohorts, namely people born in 1910-1924, 1925-1945 and 1946-1957. Figure 5 shows the birth cohort distribution of our 1503 cases.

[Figure 5 about here.]

[Table 14 about here.]

[Figure 6 about here.]

We directly consider the sequences which combine cohabitational and occupational events. The subsequences which best discriminate among cohorts are listed in Table 14. Except for (U) $\rightarrow$ (C), all displayed discriminating subsequences comprise the 'part time' (PT) event. In Figure 6, we see that there is a strong increase in the frequencies of those subsequences. We thus learn that, among the considered events, the emergence of part-time working, and especially starting to work part time after a union or after the first

childbirth, is the most prominent change which occurred over birth cohorts. The increasing frequencies of (U) $\rightarrow$ (C), is also a prominent evolution across cohorts. The latter result reflects the shift from frequent simultaneous U and C events to situations where the first childbirth occurs a year or more after moving in with a partner. This is confirmed by, for instance, the subsequence (2P) $\rightarrow$ (LH,U,C) which is the 22nd most discriminating one (result not displayed) and whose proportion falls from 27% in the oldest cohort to 12% in the youngest cohort.

6 Clustering event sequences

Besides exploring the most frequent and most discriminant subsequences, it may also be of interest to run dissimilarity-based analyses—clustering or principal coordinate analysis, for instance—of event sequences as it is typically done with state sequences (see for example Abbott and Tsay, 2000; Aisenbrey and Fasang, 2010). The only requirement for that is that the dissimilarity between time-stamped event sequences can be measured.

One possibility is to compute dissimilarities between event time stamped sequences with the OME distance, a variant of optimal-matching applicable to event sequences (Studer et al., 2010). OME is an edit distance like the optimal-matching distance between state sequences; i.e., OME is defined as the minimal cost of transforming one sequence into the other. We retain here the OME distance described in Studer et al. (2010), which extends a

proposition by Moen (2000) to the case of possible simultaneous events. The transformation operations considered by OME are

- the insertion/deletion of an event;
- a change in the time stamp of a given event;

Event dependent costs can be specified both for the insertion/deletion of an event as well as for a one-unit change in the time stamp of the event. The version implemented in TraMineR also allows to chose between absolute or relative time alignment for the pairwise comparison of sequences.

As defined above, the distance mainly depends on the mismatches—it is the cost of transforming mismatches into matches—but does not account for the total number of events in the sequences and consequently does not account for the number of existing matches. However, two sequences distant by for example 2 look very dissimilar if they contain each only two or three events, while they are very similar if they have each hundred of events. It is useful, therefore, to normalize the distance to account for the number of events in the sequence. We retain the following normalization

$$d_{N,ome}(x,y) = \frac{2d_{ome}(x,y)}{\Omega(x) + \Omega(y) + d_{ome}(x,y)}$$

where $d_{ome}(x,y)$ is the OME dissimilarity between the time-stamped event sequences x and y, and $\Omega(x)$ the total cost for inserting all the events of x. Unlike the ratio $d_{ome}(x,y)/(\Omega(x) + \Omega(y))$ proposed by Moen (2000), the retained normalized distance $d_{N,ome}$ satisfies the triangle inequality.

6.1 Types of cohabitational trajectories

To illustrate, we compute the normalized OME dissimilarity matrix of our cohabitational trajectories using a constant indel cost of 1, a constant time change cost of 0.1, and absolute time for event alignment. We do not show the 1503×1503 dissimilarity matrix which has not much interest per se, but use it to run a cluster analysis. We cluster the event sequences into five groups by partitioning them around medoids (Kaufman and Rousseeuw, 2005).[4] The resulting clusters are visualized in Figure 7. The top part renders the event sequencing, while the bottom part informs about the timing of events by displaying the survival curves until the first occurrence of the four main events, namely 'leaving home' (LH), 'starting to live alone or with friends' (A), 'moving in with a partner' (U) and 'starting to live with a child' (C). The clusters are sorted in decreasing order of frequencies and are labeled with their medoid, i.e., the sequence with minimal sum of distances to the other sequences in the cluster.

[Table 15 about here.]

[Table 16 about here.]

[Figure 7 about here.]

Unsurprisingly, we get a different cluster for each of the four typical trajectories that we could observe in Figure 2. This confirms that the OME

[4]The retained number of groups corresponds to an elbow from which quality measures such as the average silhouette (ASW = .24) and the discrepancy reduction (R^2 = .39) only grow slowly when we increase the number of clusters.

distance accounts for the sequencing of events. The survival curves show, however, that the clusters also differ in the timing of the events and in the final probability to experience the events. The fifth group corresponds to people who were living with their parents at 20 years old and did not experience any other events. Such trajectories are badly rendered by the event-sequence plots.

The first type of cohabitational trajectories (25.7%) consists in leaving home to live alone or with friends (about 5 years) before moving in with a partner, and (about 3 years) later have the first childbirth. The three events LH, U and C are spaced in time. Although this trajectory is representative of the first cluster, it does not mean that all individuals in the cluster followed exactly such a trajectory. It just means that the individuals in that cluster experienced a similar sequencing. The greyed lines in the top plot render the within-cluster diversity. We also learn from the survival curves that almost everyone in that cluster left home and experienced living alone or with a friend, that about 15% did not move in with a partner, and about 35% did not experience a childbirth.

In the second group (25.5%), the trajectories cluster around a situation where the individuals leave home, move in together with a partner and have the first childbirth simultaneously, i.e., all the same year. This is confirmed by the overlapping of the three survival curves. Surprisingly, the union seems to precede leaving home for this second cluster. However, this is essentially a consequence of the retained coding, U serving as a start event for those

who moved in with a partner before they were 20 years old, while we do not report the LH event of those persons.

In the third cluster (24.6%), we have trajectories of people who leave home to move in with a partner the same year, and (about 4 years) later have their first childbirth. The simultaneity of LH and U in this group is confirmed by the overlapping of the corresponding survival curves. About 20% of the people belonging to the cluster did not experience a childbirth.

The fourth cluster (18.6%) groups mainly people living alone when aged 20 years and who experienced spaced union start and childbirth events, the latter occurring about four years after moving in with the partner. A non negligible proportion (30%) of the individuals in the cluster did not live alone, however. From the upper plot, we can see that these are trajectories with the O start event which correspond most probably to people who lived in an institution when aged 20.

The last cluster (5.6%) is formed by those who stay unmarried with their parents during the considered ages.

From the survival curves, we learn that while almost everybody in the second and third clusters moved in with a partner, an non negligible proportion of the members of the first and fourth clusters did not experience living with a partner. The survival curves also reveal that while more than 80% of the cases experienced a childbirth in the second group, this proportion is lower in the other three clusters.

The relationship between the clusters and demographic covariates is informative. Cross tabulating cluster membership with sex, we see (Table 15) that both sexes are almost equally represented in the second and third group, while the spaced-life-events type (group 1) is clearly dominated by men. Although cluster 1 and 4 look dissimilar at first glance, a closer look reveals that the difference between them is essentially a matter of timing, the union and childbirth occurring about 3 years earlier in group 4 than in group 1. Table 15 reveals that the 'early spaced leaving home - union' cluster 4 is dominated by women while the 'late spaced leaving home - union' group 1 mainly comprises men.

The cross tabulation with birth cohort (Table 16) reveals that the proportion of individuals following a trajectory of the third type ((2P) $\stackrel{4}{\rightarrow}$ (LH,U) $\stackrel{4}{\rightarrow}$ (C) $\stackrel{18}{\longrightarrow}$) remains stable, while the proportion of the second type ((2P) $\stackrel{6}{\rightarrow}$ (C,LH,U) $\stackrel{20}{\longrightarrow}$) decreases from 38% among the older cohort to 20% for the younger cohort. In contrast, the proportion of the spaced-event types increases from 18% to 28% for cluster 1 (late-spaced events), and from 11% to 21% for cluster 4 (early-spaced events). The latter two types thus correspond to more modern ways of life organization. Finally, we notice that while staying with their parents during the active life was quite common for the oldest cohort (11%), this way of life tends to disappear in the youngest cohort (3%).

6.2 Types of occupational trajectories

Partitioning around medoids from the OME distances between occupational event sequences, we identify the 5 clusters visualized in Figure 8.[5]

[Table 17 about here.]

[Table 18 about here.]

[Figure 8 about here.]

The five identified types are in the retained order of decreased frequencies: 'Full time' (39%), 'At home after working full time' (19%), 'Short education' (18%), 'Staying at home' (12%) and 'Long education' (12%). From the survival curves, we see that all clusters also contain about 20% of individuals who experienced part-time working, the higher percentage of part time (about 45%) being in the 2nd cluster.

The correspondence with the trajectories highlighted in Figure 2 is, here, less evident than for the cohabitational trajectories. First, the most important cluster (39%) includes essentially only full-time trajectories, and the frequency of this full-time trajectory—with no other event than the starting event—is rendered by the green rectangle with regard to FT at position 1 in Figure 2. Likewise, the 'Staying at home' trajectories are represented by the hardly visible colored square with regard to AH at position 1. This rendering emphasizes the trajectory much less than the heavy lines linking successive events. A second reason for the less clear

[5] Again, the retained number of groups corresponds to an elbow in the evolution of quality measures such as the average silhouette (ASW = .39) and the discrepancy reduction ($R^2 = .55$).

correspondence between clusters and characteristic orderings is that two clusters, the 3rd and the 5th ones, are characterized by the same sequence (ED) $\rightarrow$ (FT), the difference between the two being the timing of the events for which the plot in Figure 2 does not account.

Tables 17 and 18 provide insights on the cluster composition in terms of sex and birth cohort. They reveal that the 'Full time' and 'Long education' trajectories are principally male ones, while the 'At home after working full time' type is typically female. The type 'At home', also dominated by women, declines across birth cohorts in favor of the 'Short education', (ED) $\stackrel{1}{\rightarrow}$ (FT), and 'At home after working full time', (FT) $\stackrel{6}{\rightarrow}$ (AH), types of trajectory.

7 Conclusion

Three methods for discovering useful knowledge from life event histories have been addressed: the mining of frequent sequential patterns, the identification of the most discriminant subsequences between given groups, and the definition of types of trajectories by means of unsupervised clustering. As illustrated by their application on a Swiss dataset on cohabitational and occupational trajectories, the three approaches provide complementary insights on the way Swiss people organized their life course during the 20th century and how it changed over time. The mining of frequent sequential patterns permits to find out the overall most common characteristics of the analyzed sequences. The search for discriminant subsequences helps to understand the salient distinctions between groups such as between men or women or between successive birth cohorts. Finally, clustering and the resulting suggested typology serves to identify dominant types of trajectories.

The mining of frequent sequential patterns has received a lot of attention from the data mining community which proposed plenty of efficient algorithms, the main concern being the scalability to very large datasets and the ability to handle time and content constraints. The mining of frequent patterns from life course data is quite new, however. The only such application that we know about is the study by Blockeel et al. (2001). One reason for this apparently low interest for the mining of frequent sequential patterns in social sciences is that the available tools remained hardly accessible to non-computer scientists. We put, therefore, much effort in implementing simple commands—illustrated in the text—in our TraMineR R-package to assist the user in putting the data in a suitable form and in running the mining process under various constraints. The tools for finding discriminant subsequences and computing OME distances between event sequences are unique features of TraMineR and are currently not available elsewhere.

It is worth mentioning that the clustering of event sequences as illustrated in Section 6, is just one possible use of the pairwise OME dissimilarities. Indeed, once we have the OME distances, we can run any dissimilarity-based analysis including, among others, self-organizing maps (Kohonen, 1997; Massoni et al., 2009), principal coordinate analysis (Gower, 1966), the search for representative sequences (Gabadinho et al., 2011), ANOVA-like discrepancy analysis and regression trees (Studer et al., 2011).

What did we learn from our exploration of Swiss life courses? As expected, the findings from frequent and discriminant subsequences essentially concern the sequencing of events. The main result is that while experiencing simultaneously on the same year leaving home, moving in together with a partner and the first

childbirth predominated among the older cohorts (people born between 1910 and 1924), the norm tends towards more spaced events in younger cohorts. Differences between male and female cohabitational trajectories are not very important and essentially concern event timing. Regarding occupational events, however, there are very strong differences. The results clearly demonstrate that the 20th century is characterized by the emergence of part-time working and that this phenomenon mainly concerns women. While staying at home remains quite common for women, we observed a slow shift across cohorts from cases where women stay at home since 20 years old to a model where women first work full time during their early 20's and only later stop working to stay at home, generally after the first childbirth. The cluster analysis confirmed those results but complemented them by emphasizing important timing differences. For example, women most often stop education to start working full time around 21 years old while it is more typical for men to do so around the age of 26.

Acknowledgements

This publication results from research work executed within the framework of the Swiss National Centre of Competence in Research LIVES, which is financed by the Swiss National Science Foundation. The authors are grateful to the Swiss National Science Foundation for its financial support. In addition the authors also thank Nicolas S. Müller for his participation in the coding of the TraMineR functions for event sequences.

References

Abbott, A. (1983). Sequences of social events: Concepts and methods for the analysis of order in social processes. *Historical Methods 16*(4), 129–147.

Abbott, A. (1991). The order of professionalization: An empirical analysis. *Work and Occupations 18*(4), 355–384.

Abbott, A. and J. Forrest (1986). Optimal matching methods for historical sequences. *Journal of Interdisciplinary History 16*, 471–494.

Abbott, A. and A. Tsay (2000). Sequence analysis and optimal matching methods in sociology, Review and prospect. *Sociological Methods and Research 29*(1), 3–33. (With discussion, pp 34-76).

Agrawal, R., H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo (1995). Fast discovery of association rules. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*, pp. 307–328. Menlo Park, CA: AAAI Press.

Agrawal, R. and R. Srikant (1994). Fast algorithm for mining association rules in large databases. In J. B. Bocca, M. Jarke, and C. Zaniolo (Eds.), *Proceedings 1994 International Conference on Very Large Data Base (VLDB'94), Santiago de Chile*, San-Mateo, pp. 487–499. Morgan-Kaufman.

Agrawal, R. and R. Srikant (1995). Mining sequential patterns. In P. S. Yu and A. L. P. Chen (Eds.), *Proceedings of the International Conference on Data Engeneering (ICDE), Taipei, Taiwan*, pp. 487–499. IEEE Computer Society.

Aisenbrey, S. and A. E. Fasang (2010). New life for old ideas: The "second wave" of sequence analysis bringing the "course" back into the life course. *Sociological Methods and Research 38*(3), 430–462.

Bettini, C., X. S. Wang, and S. Jajodia (1996). Testing complex temporal relationships involving multiple granularities and its application to data mining (extended abstract). In *PODS '96: Proceedings of the fifteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, New York, pp. 68–78. ACM Press.

Billari, F. C., J. Fürnkranz, and A. Prskawetz (2006). Timing, sequencing, and quantum of life course events: A machine learning approach. *European Journal of Population 22*(1), 37–65.

Blockeel, H., J. Fürnkranz, A. Prskawetz, and F. Billari (2001). Detecting temporal change in event sequences: An application to demographic data. In L. De Raedt and A. Siebes (Eds.), *Principles of Data Mining and Knowledge Discovery: 5th European Conference, PKDD 2001*, Volume LNCS 2168, pp. 29–41. Freiburg in Brisgau: Springer.

Blossfeld, H.-P., K. Golsch, and G. Rohwer (2007). *Event History Analysis with Stata*. Mahwah NJ: Lawrence Erlbaum.

Brzinsky-Fay, C., U. Kohler, and M. Luniak (2006). Sequence analysis with Stata. *The Stata Journal 6*(4), 435–460.

Bürgin, R., G. Ritschard, and E. Rousseaux (2012). Visualisation de séquences

d'événements. *Revue des Nouvelles Technologies de l'Information (RNTI) E-23*, 559–560.

Elzinga, C. H. (2003). Sequence similarity: A non-aligning technique. *Sociological Methods and Research 31*, 214–231.

Elzinga, C. H. (2010). Complexity of categorical time series. *Sociological Methods & Research 38*(3), 463–481.

Gabadinho, A., G. Ritschard, N. S. Müller, and M. Studer (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software 40*(4), 1–37.

Gabadinho, A., G. Ritschard, M. Studer, and N. S. Müller (2009). Mining sequence data in R with the TraMineR package: A user's guide. Technical report, Department of Econometrics and Laboratory of Demography, University of Geneva, Geneva.

Gabadinho, A., G. Ritschard, M. Studer, and N. S. Müller (2011). Extracting and rendering representative sequences. In A. Fred, J. L. G. Dietz, K. Liu, and J. Filipe (Eds.), *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Volume 128 of *Communications in Computer and Information Science (CCIS)*, pp. 94–106. Springer-Verlag.

Ghisletta, P. (2012). Recursive partitioning to study terminal decline in the Berlin aging study. In J. J. McArdle and G. Ritschard (Eds.), *Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences*, Quantitative Methodology. New York: Routledge.

Goodman, L. A. and W. H. Kruskal (1954). Measures of association for cross classifications. *Journal of the American Statistical Association 49*, 732–764.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika 53*(3/4), 325–338.

Joshi, M. V., G. Karypis, and V. Kumar (2001). A universal formulation of sequential patterns. In *Proceedings of the KDD'2001 workshop on Temporal Data Mining, San Fransisco, August 2001*.

Kaufman, L. and P. J. Rousseeuw (2005). *Finding Groups in Data*. Hoboken: John Wiley & Sons.

Kohonen, T. (1997). *Self-Organizing Maps* (2nd ed.). in Information Sciences. Heidelberg: Springer.

Mannila, H., H. Toivonen, and A. I. Verkamo (1995). Discovering frequent episodes in sequences. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada, August 20-21, 1995*, pp. 210–215. AAAI Press.

Mannila, H., H. Toivonen, and A. I. Verkamo (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery 1*(3), 259–289.

Masseglia, F. (2002). *Algorithmes et applications pour l'extraction de motifs séquentiels dans le domaine de la fouille de données : de l'incrémental au temps réel*. Ph. D. thesis, Université de Versailles Saint-Quentin en Yvelines.

Massoni, S., M. Olteanu, and P. Rousset (2009). Career-path analysis using optimal matching and self-organizing maps. In J. C. Príncipe and R. Miikkulainen (Eds.), *Advances in Self-Organizing Maps: 7th International Workshop, WSOM 2009, St. Augustine, FL, USA, June 8-10, 2009*, Volume 5629 of *Lecture Notes in Computer Science*, pp. 154–162. Berlin: Springer.

Moen, P. (2000). *Attribute, Event Sequence, and Event Type Similarity Notions for Data Mining*. PhD thesis, University of Helsinki.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Ritschard, G., A. Gabadinho, N. S. Müller, and M. Studer (2008). Mining event histories: A social science perspective. *International Journal of Data Mining, Modelling and Management 1*(1), 68–90.

Ritschard, G., A. Gabadinho, M. Studer, and N. S. Müller (2009). Converting between various sequence representations. In Z. Ras and A. Dardzinska (Eds.), *Advances in Data Management*, Volume 223 of *Studies in Computational Intelligence*, pp. 155–175. Berlin: Springer-Verlag.

Scott, S. B., B. R. Whitehead, C. S. Bergeman, and L. Pitzer (2012). Understanding global perceptions of stress in adulthood through tree-based exploratory data mining. In J. J. McArdle and G. Ritschard (Eds.), *Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences*, Quantitative Methodology. New York: Routledge.

Srikant, R. and R. Agrawal (1996). Mining sequential patterns: Generalizations and performance improvements. In P. M. G. Apers, M. Bouzeghoub, and G. Gardarin (Eds.), *Advances in Database Technologies – 5th International Conference on Extending Database Technology (EDBT'96), Avignon, France*, Volume 1057, pp. 3–17. Springer-Verlag.

Studer, M., N. S. Müller, G. Ritschard, et A. Gabadinho (2010). Classer, discriminer et visualiser des séquences d'événements. *Revue des nouvelles technologies de l'information RNTI E-19*, 37–48.

Studer, M., G. Ritschard, A. Gabadinho, et N. S. Müller (2011). Discrepancy analysis of state sequences. *Sociological Methods and Research 40*(3), 471–510.

Theil, H. (1970). On the estimation of relationships involving qualitative variables. *American Journal of Sociology 76*, 103–154.

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

Yamaguchi, K. (1991). *Event history analysis*. ASRM 28. Newbury Park and London: Sage.

Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning 42*(1/2), 31–60.

Zhou, Y., K. M. Kadlec, and J. J. McArdle (2012). Predicting mortality from demographics and specific cognitive abilities in the Hawaii family study of cognition. In J. J. McArdle and G. Ritschard (Eds.), *Contemporary Issues in Exploratory*

Data Mining in the Behavioral Sciences, Quantitative Methodology. New York: Routledge.

List of Figures

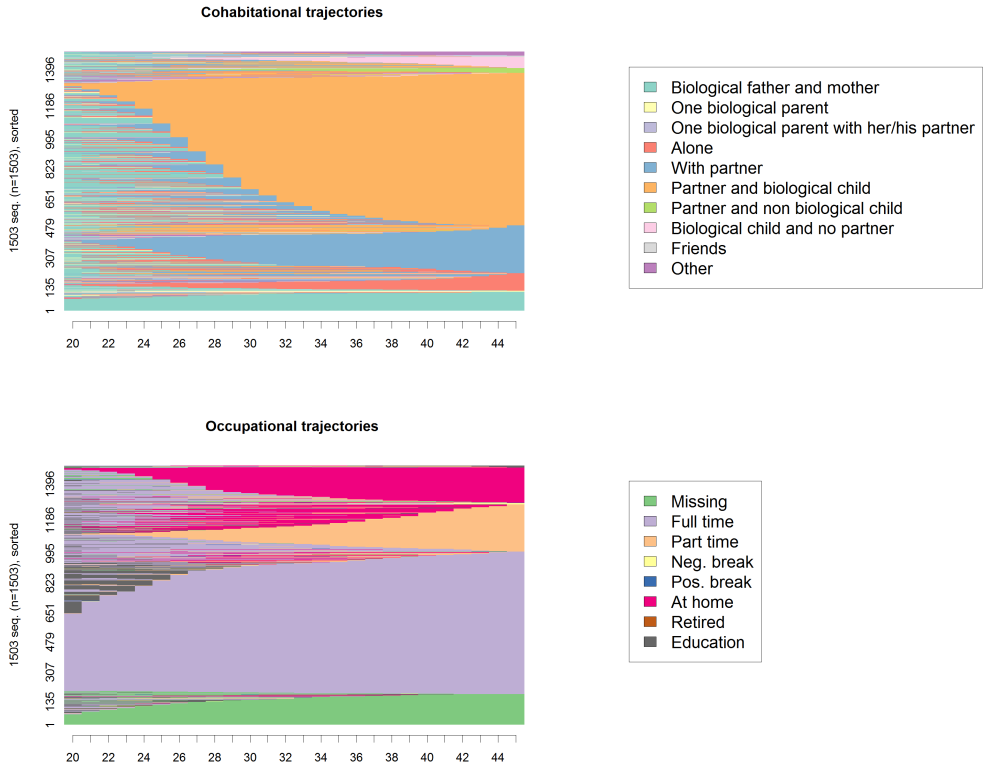

Figure 1: Cohabitational and Occupational State Sequences

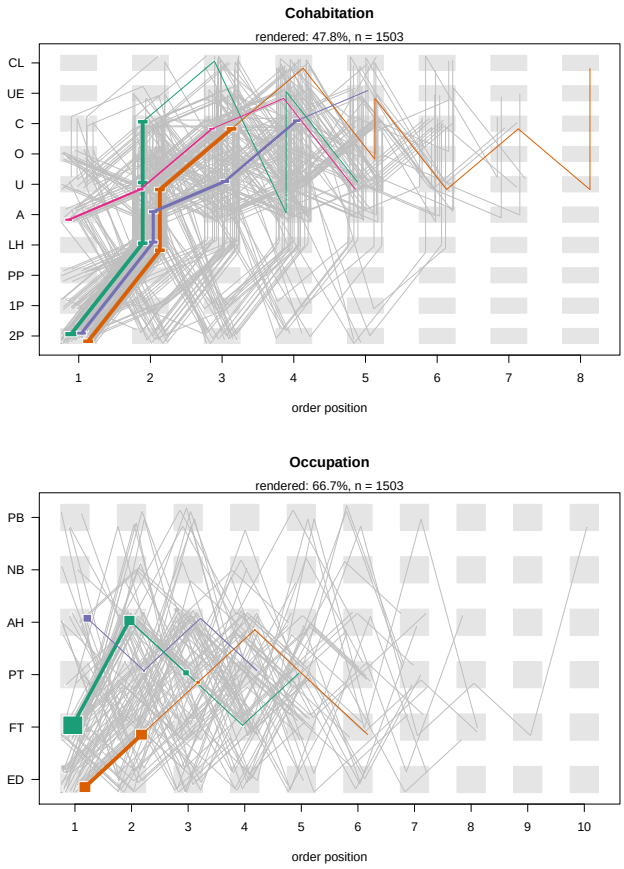

Figure 2: Event Sequencing in cohabitational and occupational trajectories

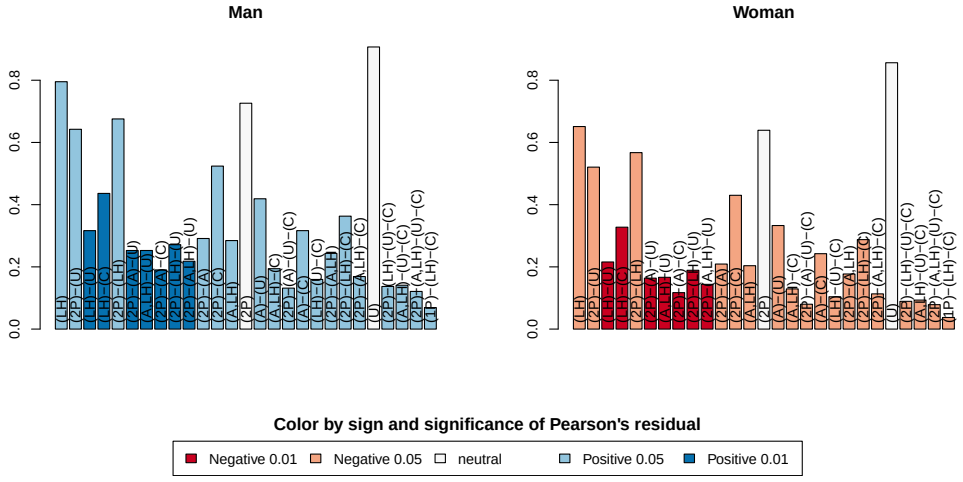

Figure 3: Cohabitational subsequences that discriminate sex at the 1% level

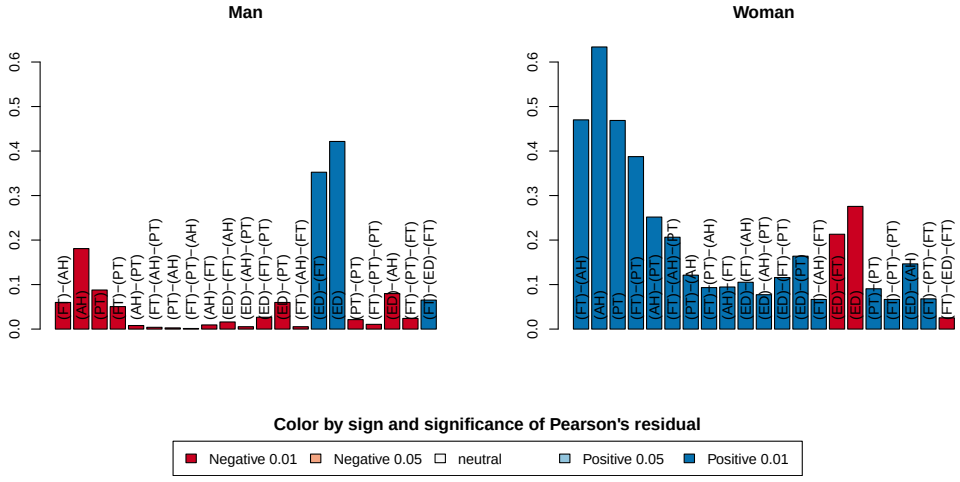

Figure 4: Occupational subsequences that discriminate sex at the 0.1% level

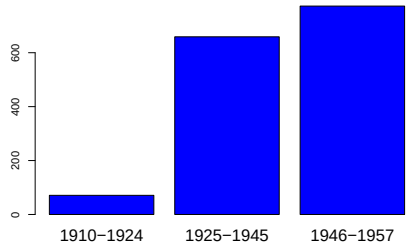

Figure 5: Birth cohort distribution

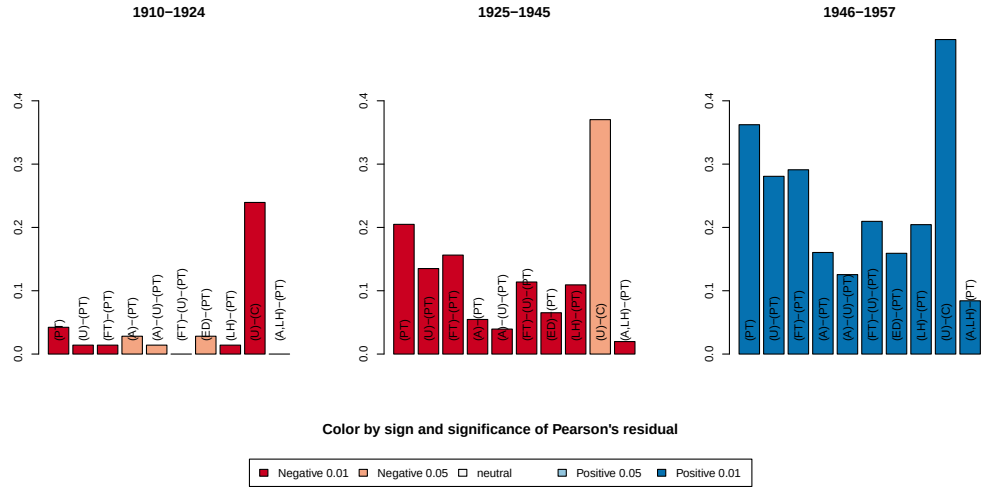

Figure 6: Mixed events: Subsequences that best discriminate birth cohorts

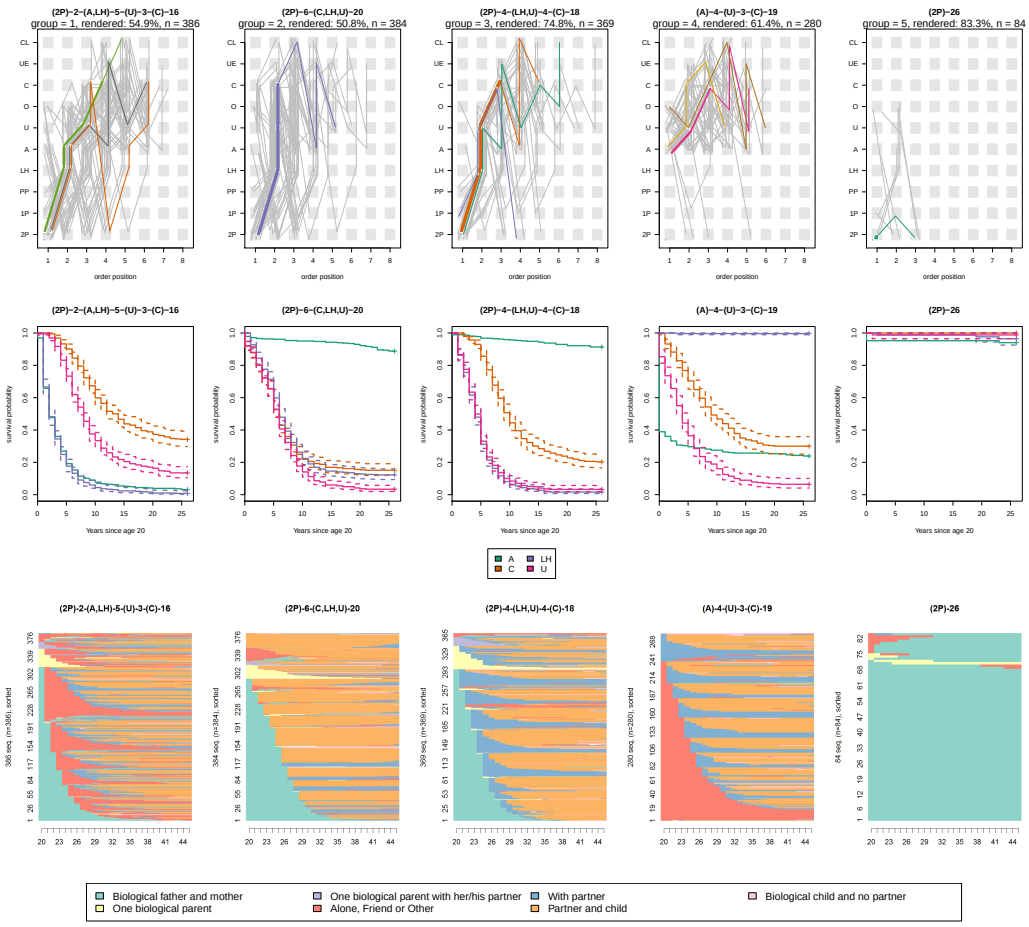

Figure 7: Cohabitational trajectories clustered from dissimilarities between event sequences

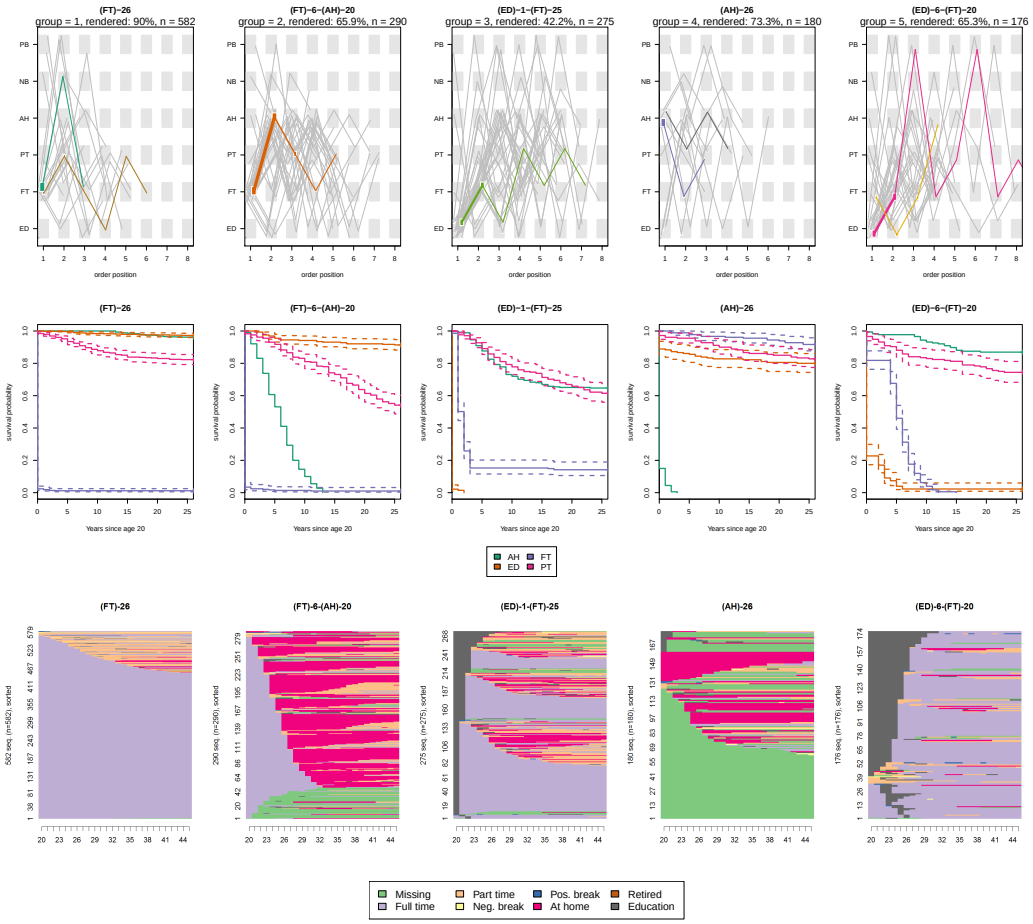

Figure 8: Occupational trajectories clustered from dissimilarities between event sequences

List of Tables

Table 1: Short and long state labels

Cohabitational		Occupational	
2P	Biological father and mother	Mi	Missing
1P	One biological parent	FT	Full time
PP	One biological parent with her/his partner	PT	Part time
A	Alone	NB	Neg. break
U	With partner	PB	Pos. break
UC	Partner and biological child	AH	At home
UN	Partner and non biological child	RE	Retired
C	Biological child and no partner	ED	Education
F	Friends		
O	Other		

Table 2: Cohabitational events

Short	Long label
2P	Start living with both parents
1P	Start living with one parent
PP	Start living with one parent and her/his partner
LH	Leaving home
A	Start living alone
U	Move in together with partner
UE	Separation
C	Start living with a child
CL	Last child leaves home
O	Start other living arrangement

Table 3: Events associated to cohabitational state transitions

	2P	1P	PP	A	U	UC	UN	C	F	O
2P	"2P"	"1P"	"PP"	"LH,A"	"LH,U"	"LH,U,C"	"LH,U,C"	"LH,C"	"LH,A"	"LH,O"
1P	"2P"	"1P"	"PP"	"LH,A"	"LH,U"	"LH,U,C"	"LH,U,C"	"LH,C"	"LH,A"	"LH,O"
PP	"2P"	"1P"	"PP"	"LH,A"	"LH,U"	"LH,U,C"	"LH,U,C"	"LH,C"	"LH,A"	"LH,O"
A	"2P"	"1P"	"PP"	"A"	"U"	"U,C"	"U,C"	"C"	""	"O"
U	"2P"	"1P"	"PP"	"UE,A"	"U"	"C"	"C"	"C"	"UE,A"	"UE,O"
UC	"2P"	"1P"	"PP"	"UE,CL,A"	"CL"	"U,C"	"CL,C"	"UE"	"UE,CL,A"	"UE,CL,O"
UN	"2P"	"1P"	"PP"	"UE,CL,A"	"CL"	"C"	"U,C"	"UE,C"	"UE,CL,A"	"UE,CL,O"
C	"2P"	"1P"	"PP"	"CL,A"	"CL,U"	"U"	"CL,C"	"C"	"CL,A"	"CL,O"
F	"2P"	"1P"	"PP"	""	"U"	"U,C"	"U,C"	"C"	"A"	"O"
O	"2P"	"1P"	"PP"	"A"	"U"	"U,C"	"U,C"	"C"	"A"	"O"

Table 4: Example of vertical time-stamped event data

Id	Time stamp	Event
101	22	LHome
101	22	Union
101	25	Marriage
101	27	Childbirth
102	18	Union
102	..	...

Table 5: First 5 cohabitational event sequences

	sequence
1	(2P) $\xrightarrow{1}$ (LH,U) $\xrightarrow{25}$
2	(2P) $\xrightarrow{2}$ (LH,U) $\xrightarrow{6}$ (C) $\xrightarrow{18}$
3	(2P) $\xrightarrow{1}$ (A,LH) $\xrightarrow{5}$ (U) $\xrightarrow{2}$ (C) $\xrightarrow{18}$
4	(2P) $\xrightarrow{8}$ (LH,U) $\xrightarrow{1}$ (C) $\xrightarrow{17}$
5	(A) $\xrightarrow{6}$ (C,U) $\xrightarrow{20}$

Table 6: Most frequent cohabitational subsequences with at least 2 events

	Subsequence	Support	Count	#Transitions	#Events
1	(2P) → (LH)	0.621	934	2	2
2	(2P) → (U)	0.582	874	2	2
3	(2P) → (C)	0.477	717	2	2
4	(LH,U)	0.454	682	1	2
5	(U) → (C)	0.429	645	2	2
6	(2P) → (LH,U)	0.392	589	2	3
7	(LH) → (C)	0.382	574	2	2
8	(A) → (U)	0.376	565	2	2
9	(2P) → (LH) → (C)	0.325	489	3	3
10	(C,U)	0.291	437	1	2

Table 7: Most frequent cohabitational subsequences with at least 2 events and a 3-year maximum time span

	Subsequence	Support	Count	#Transitions	#Events
1	(LH,U)	0.454	682	1	2
2	(C,U)	0.291	437	1	2
3	(2P) → (LH)	0.275	414	2	2
4	(U) → (C)	0.274	412	2	2
5	(A,LH)	0.244	367	1	2
6	(C,LH)	0.180	270	1	2
7	(C,LH,U)	0.175	263	1	3
8	(LH) → (C)	0.166	250	2	2
9	(A) → (U)	0.158	237	2	2
10	(2P) → (A)	0.148	223	2	2

Table 8: Most frequent occupational subsequences with at least 2 transitions

	Subsequence	Support	Count	#Transitions	#Events
1	(ED) → (FT)	0.283	425	2	2
2	(FT) → (AH)	0.265	398	2	2
3	(FT) → (PT)	0.219	329	2	2
4	(AH) → (PT)	0.130	195	2	2
5	(ED) → (AH)	0.113	170	2	2
6	(ED) → (PT)	0.112	168	2	2
7	(FT) → (FT)	0.112	168	2	2
8	(FT) → (AH) → (PT)	0.105	158	3	3
9	(FT) → (ED)	0.073	109	2	2
10	(ED) → (FT) → (PT)	0.071	107	3	3

Table 9: Most frequent occupational subsequences with at least 2 transitions and a 3-year maximum time span

	Subsequence	Support	Count	#Transitions	#Events
1	(ED) → (FT)	0.185	288	2	2
2	(FT) → (AH)	0.067	100	2	2
3	(ED) → (AH)	0.042	73	2	2
4	(PT) → (FT)	0.036	56	2	2
5	(PT) → (AH)	0.034	53	2	2
6	(ED) → (PT)	0.031	52	2	2

Table 10: Most frequent subsequences of combined cohabitational-occupational sequences

	Subsequence	Support	Count	#Transitions	#Events
1	(FT) → (U)	0.695	1045	2	2
2	(2P) → (LH)	0.621	934	2	2
3	(FT) → (C)	0.583	876	2	2
4	(2P) → (U)	0.582	874	2	2
5	(FT) → (LH)	0.555	834	2	2
6	(2P) → (C)	0.477	717	2	2
7	(LH,U)	0.454	682	1	2
8	(U) → (C)	0.429	645	2	2
9	(2P) → (LH,U)	0.392	589	2	3
10	(LH) → (C)	0.382	574	2	2
11	(2P,FT)	0.378	568	1	2
12	(A) → (U)	0.376	565	2	2

Table 11: Cohabitational subsequences that best discriminate sex

	Subsequence	Chi-2	Support	Freq. Men	Freq. Women	Diff
1	(LH)	38.3	0.72	0.795	0.651	0.144
2	(2P) → (U)	22.4	0.58	0.642	0.521	0.122
3	(LH) → (U)	19.0	0.27	0.316	0.216	0.101
4	(LH) → (C)	18.3	0.38	0.436	0.328	0.109
5	(2P) → (LH)	18.3	0.62	0.676	0.567	0.108
6	(2P) → (A) → (U)	17.5	0.21	0.253	0.164	0.089

Table 12: Occupational subsequences that best discriminate sex

	Subsequence	Chi-2	Support	Freq. Men	Freq. Women	Diff
1	(FT) → (AH)	322.7	0.26	0.060	0.470	-0.410
2	(AH)	317.5	0.41	0.181	0.634	-0.453
3	(PT)	269.7	0.28	0.088	0.469	-0.381
4	(FT) → (PT)	247.5	0.22	0.051	0.387	-0.337
5	(AH) → (PT)	195.5	0.13	0.008	0.252	-0.244
6	(FT) → (AH) → (PT)	161.5	0.11	0.004	0.206	-0.202

Table 13: Mixed events: Subsequences that best discriminate sex

	Subsequence	Chi-2	Support	Freq. Men	Freq. Women	Diff
1	(FT) → (AH)	322.7	0.26	0.060	0.470	-0.410
2	(AH)	317.5	0.41	0.181	0.634	-0.453
3	(PT)	269.7	0.28	0.088	0.469	-0.381
4	(U) → (PT)	260.4	0.20	0.036	0.373	-0.337
5	(FT) → (PT)	247.5	0.22	0.051	0.387	-0.337
6	(FT) → (U) → (AH)	228.2	0.16	0.016	0.302	-0.286
7	(U) → (AH)	226.0	0.20	0.041	0.350	-0.309
8	(AH) → (PT)	195.5	0.13	0.008	0.252	-0.244
9	(C) → (PT)	193.3	0.15	0.019	0.273	-0.254
10	(FT) → (U) → (PT)	192.7	0.16	0.027	0.289	-0.262

Table 14: Mixed events: Subsequences that best discriminate birth cohorts

	Subsequence	Chi-2	Support	1910-25	1926-45	1946-57
1	(PT)	64.5	0.28	0.042	0.205	0.362
2	(U) → (PT)	63.0	0.20	0.014	0.135	0.281
3	(FT) → (PT)	56.1	0.22	0.014	0.156	0.291
4	(A) → (PT)	46.3	0.11	0.028	0.055	0.160
5	(A) → (U) → (PT)	39.4	0.08	0.014	0.039	0.125
6	(FT) → (U) → (PT)	38.5	0.16	0.000	0.114	0.210
7	(ED) → (PT)	36.8	0.11	0.028	0.065	0.159
8	(LH) → (PT)	35.9	0.15	0.014	0.109	0.204
9	(U) → (C)	34.2	0.43	0.239	0.370	0.497
10	(A,LH) → (PT)	34.0	0.05	0.000	0.020	0.084

Table 15: Cohabitational trajectory types, distribution by sex

	Man	Woman	Overall
(2P) $\xrightarrow{2}$ (A,LH) $\xrightarrow{5}$ (U) $\xrightarrow{3}$ (C) $\xrightarrow{16}$	0.298	0.216	0.257
(2P) $\xrightarrow{6}$ (C,LH,U) $\xrightarrow{20}$	0.266	0.245	0.255
(2P) $\xrightarrow{4}$ (LH,U) $\xrightarrow{4}$ (C) $\xrightarrow{18}$	0.249	0.242	0.246
(A) $\xrightarrow{4}$ (U) $\xrightarrow{3}$ (C) $\xrightarrow{19}$	0.138	0.234	0.186
(2P) $\xrightarrow{26}$	0.049	0.063	0.056

Table 16: Cohabitational trajectory types, distribution by birth cohorts

	1910-1924	1925-1945	1946-1957	Overall
(2P) $\xrightarrow{2}$ (A,LH) $\xrightarrow{5}$ (U) $\xrightarrow{3}$ (C) $\xrightarrow{16}$	0.183	0.235	0.282	0.257
(2P) $\xrightarrow{6}$ (C,LH,U) $\xrightarrow{20}$	0.380	0.310	0.198	0.255
(2P) $\xrightarrow{4}$ (LH,U) $\xrightarrow{4}$ (C) $\xrightarrow{18}$	0.211	0.211	0.278	0.246
(A) $\xrightarrow{4}$ (U) $\xrightarrow{3}$ (C) $\xrightarrow{19}$	0.113	0.164	0.212	0.186
(2P) $\xrightarrow{26}$	0.113	0.080	0.030	0.056

Table 17: Occupational trajectory types, distribution by sex

	Man	Woman	Overall
(FT) $\xrightarrow{26}$	0.488	0.286	0.387
(FT) $\xrightarrow{6}$ (AH) $\xrightarrow{20}$	0.041	0.345	0.193
(ED) $\xrightarrow{1}$ (FT) $\xrightarrow{25}$	0.185	0.181	0.183
(AH) $\xrightarrow{26}$	0.100	0.140	0.120
(ED) $\xrightarrow{6}$ (FT) $\xrightarrow{20}$	0.186	0.048	0.117

Table 18: Occupational trajectory types, distribution by birth cohort

	1910-1924	1925-1945	1946-1957	Overall
(FT) $\xrightarrow{26}$	0.338	0.404	0.378	0.387
(FT) $\xrightarrow{6}$ (AH) $\xrightarrow{20}$	0.141	0.209	0.184	0.193
(ED) $\xrightarrow{1}$ (FT) $\xrightarrow{25}$	0.127	0.155	0.212	0.183
(AH) $\xrightarrow{26}$	0.239	0.135	0.096	0.120
(ED) $\xrightarrow{6}$ (FT) $\xrightarrow{20}$	0.155	0.097	0.131	0.117