Searching for typical life trajectories applied to childbirth histories*

Alexis Gabadinho and Gilbert Ritschard**

NCCR LIVES and Institute for Demographic and Life Course Studies
University of Geneva, 40, bd du Pont-d'Arve, CH-1211 Geneva, Switzerland
{alexis.gabadinho,gilbert.ritschard}@unige.ch
http://mephisto.unige.ch/TraMineR

Abstract. We address, in this chapter, the identification of typical patterns that best characterize a set of sequences. More specifically, we focus on data-driven methods that search for the typical patterns among the observed sequences. In life course studies, such typical sequences serve, for instance, to describe ideal-type life trajectories; i.e., the common way(s) of organizing our life. We propose an heuristic for extracting an as small as possible subset of patterns that covers a given percentage of all sequences. The method is based on the concept of neighborhood, the coverage being the number of representative sequences in the neighborhood of the retained representative patterns. We illustrate the scope of the method by applying it to the study of childbirth histories of Swiss women. The representative sets obtained for six successive birth cohorts clearly exhibit that while patterns with three or four childbirths were common for women of older cohorts, it is no longer the case for younger cohorts. At the same time, representative plots permit to identify the typical timing of the successive childbirths.

Key words: Categorical Sequences, Representative patterns, Ideal-type trajectories, Medoid, Centrality, Neighborhood density.

1 Introduction

The analysis of categorical sequences has gained much importance in life course analysis since Abbott (Abbott and Forrest, 1986; Abbott and Hrycak, 1990) popularized optimal-matching methods in the social sciences. The success of sequence analysis for studying life trajectories such as professional careers or familial life courses is largely attributable to its holistic perspective, which, by considering each sequence as a whole, usefully complements approaches such as event history analysis that focus on a given event (Billari, 2005).

* Cite as Gabadinho, A. and G. Ritschard (2013), Searching for typical life trajectories applied to childbirth histories. In R. Levy and E. Widmer (Eds), *Gendered life courses - Between individualization and standardization. A European approach applied to Switzerland*, Vienna: LIT, pp. 287-312.

** This work results from research done in the framework of the NCCR LIVES, financed by the Swiss National Science Foundation (see http://www.lives-nccr.ch).

In sequence analysis, the analyst is typically interested in comparing groups of sequences defined either by a clustering procedure or by the values of a covariate such as sex, birth cohort or socio-economic level. It is crucial for this comparison to get a clear idea of what are the characteristic patterns of each group. Finding out those characteristics is difficult, however, due to the usually high number of distinct patterns present in each of the groups. The sequence of transversal state distributions gives some kind of average view, but it does not render individual patterns and hence says nothing about their diversity. Likewise a central sequence pattern does not inform about the diversity and may be quite far from the other sequences in the group when this diversity is high. The issue of summarizing efficiently a group of sequences is still unresolved (Aassve et al., 2007). A common definition of what a "typical sequence" is and an approach for characterizing typical sequences are lacking in the literature.

Even in the most recent works, the identification of typical sequences is often simply done through visual inspection (Abbott and Hrycak, 1990; Wiggins et al., 2007; Martin et al., 2008). This approach has some serious drawbacks. It is subjective and becomes hardly practicable when the number of sequences and hence the number of different patterns increases. Automated data driven solutions, such as the most centrally located sequence considered by Abbott and Hrycak (1990, page 164) or the sequence of modal states proposed in Aassve et al. (2007), essentially consist in determining a single representative sequence. The most central sequence is searched among the observed sequences while the modal state is an artificial construction that may possibly not be observed. This suggests distinguishing between two approaches for characterizing representative sequences: Searching representatives among the observed sequences, and building artificial synthetic sequences that would satisfy some criteria such as minimizing the sum of distances to the sequences in the set. The main problem with the latter approach is that the optimization process does by no way ensure the plausibility of the solution. For example, we could end up with a representative in which a state 'married' is followed by a state 'single', which is indeed not possible in an individual sequence. We therefore focus in this paper on the first approach; that is, we look for representatives among the observed sequences. We propose a general framework to select typical representative sequences according to different representativeness criteria. We also develop a series of measures of the representativeness quality of the selected sequences. Some of those measures such as the coverage—percentage of sequences lying in a given neighborhood of the representatives—will prove useful to monitor and control the size of the representative set and/or the quantity of information it carries.

We illustrate our presentation by means of a demographic issue, namely the evolution of the fertility histories of Swiss women. We consider data for women born before 1950 stemming from the 2000 Swiss federal census, from which we build sequences describing the childbirth histories of six successive five-year-long birth cohorts.

The chapter is organized as follows. In Section 2, we present the illustrative application framework with a preliminary descriptive analysis of the considered

childbirth histories. In Section 3 we recall basic concepts of sequence rendering and analysis. We address the concept of typicality in Section 4 where we also discuss the scope and limits of existing data-driven approaches. We describe our approach for searching for a set of representatives in Section 5 and apply it to the study of Swiss childbirth histories in Section 7. Finally, we make a few concluding remarks in 8.

2 Illustrative Data

The application retained for illustrating the proposed methods is an original analysis of the evolution of Swiss childbirth histories over birth cohorts. It is based on data of the 2000 Swiss federal census. We derived the analyzed fertility histories from questions about the birth year of the first four children of each respondent. We focus on completed histories only and we therefore consider only women aged 50 years at census time. We discarded the few women born before 1920 and women with missing information regarding the birth date of one of their children. We organized the remaining 502386 cases into six 5-year-long cohorts.[1]

Eventually, we extracted from each cohort a random sample of each 1500 women, yielding a tractable total of 9000 sequences for the analysis. The retained data are summarized in Table 1, where P0, . . . , P4+ stand for the completed parity—the total number of childbirths—at the end—50 years—of the fertility life. P4+ means 4 or more children.

Table 1: Number of cases, sample size, distinct patterns and completed parity.

	N	Sample	Patterns	P0	P1	P2	P3	P4+
1920-24	61384	1500	650	22.1	14.7	25.5	17.9	19.9
1925-29	70217	1500	649	19.1	14.8	26.7	19.1	20.2
1930-34	76992	1500	654	16.1	14.6	28.2	23.0	18.1
1935-39	82167	1500	617	15.6	13.5	32.5	22.2	16.2
1940-44	98418	1500	517	14.6	17.5	39.9	18.8	9.3
1945-49	113208	1500	469	18.6	15.9	43.3	17.1	5.1
1920-49	502386	9000	1870	17.7	15.2	32.7	19.7	14.8

2.1 Preliminary descriptive analysis

The original data is in the form of event histories; that is, a series of time stamped events—the birth years of the children. In demography, such longitudinal data is

[1] Since we have retrospective data, we measure the fertility histories of the survivors only. This raises, given the possible link between the probability of surviving in 2000 and the number of childbirths, the question of the representativeness of the oldest cohorts.

Table 2: Quartiles for age at first and second childbirths and interval between them, by birth cohort and completed fertility.

cohort	comp	n1_q1	n1_q2	n1_q3	n2_q1	n2_q2	n2_q3	n1_2_q1	n1_2_q2	n1_2_q3
1920-24	P0									
1920-24	P1	25.0	28.0	32.0						
1920-24	P2	24.0	27.0	30.0	28.0	31.0	35.0	2.0	3.0	5.0
1920-24	P3	23.0	26.0	28.0	26.0	28.0	31.0	2.0	2.0	3.0
1920-24	P4+	22.0	25.0	27.0	25.0	27.0	29.0	1.0	2.0	2.0
1925-29	P0									
1925-29	P1	24.0	29.0	34.0						
1925-29	P2	24.0	27.0	30.0	27.0	31.0	34.0	2.0	3.0	5.0
1925-29	P3	23.0	25.0	28.0	26.0	28.0	31.0	2.0	2.0	3.0
1925-29	P4+	23.0	25.0	27.0	25.0	27.0	29.0	1.0	2.0	2.0
1930-34	P0									
1930-34	P1	24.0	28.0	33.5						
1930-34	P2	24.0	26.0	29.0	27.0	30.0	33.0	2.0	3.0	4.0
1930-34	P3	23.0	25.0	27.0	25.0	27.0	30.0	2.0	2.0	3.0
1930-34	P4+	22.0	24.0	26.0	24.0	26.0	29.0	1.0	2.0	2.0
1935-39	P0									
1935-39	P1	24.0	29.0	32.0						
1935-39	P2	23.0	25.0	28.0	27.0	29.0	32.0	2.0	3.0	5.0
1935-39	P3	22.0	24.0	26.0	25.0	26.0	29.0	2.0	2.0	3.0
1935-39	P4+	21.0	23.0	25.0	23.0	25.0	27.0	1.0	2.0	2.0
1940-44	P0									
1940-44	P1	23.0	26.0	30.0						
1940-44	P2	23.0	25.0	28.0	26.0	29.0	32.0	2.0	3.0	4.0
1940-44	P3	22.0	24.0	26.0	24.0	26.0	29.0	1.0	2.0	3.0
1940-44	P4+	21.0	23.0	25.0	23.0	25.0	27.0	1.0	2.0	2.0
1945-49	P0									
1945-49	P1	23.0	26.0	30.5						
1945-49	P2	22.0	24.0	27.2	25.0	28.0	31.0	2.0	3.0	4.0
1945-49	P3	21.0	23.5	27.0	24.0	27.0	30.0	2.0	2.0	3.0
1945-49	P4+	20.0	22.0	24.0	22.0	24.0	26.0	1.0	2.0	2.8

typically analysed by calculating age specific or total fertility rates, descriptive statistics of the timing and of the number of births. For example, in their study of the postwar fertility patterns in the Federal Republic of Germany, Tuma and Huinink (1990) discuss, among others, quartiles of the age at childbirths of given ranks and selected childbirth intervals—duration between two successive childbirths—by sex and cohort. They also comment the completed fertility of the cohorts. We give similar descriptive statistics for our data in Table 2.

Event history approaches are also widely used in population studies (for example in Schoumaker and Hayford, 2004). It includes life tables, parametric and non-parametric estimation of survival curves and, with a more causal perspective, regression-like methods—Cox model, logistic regression on person-period

data, Poisson regression and other models for survival data (see for instance Yamaguchi, 1991). Event history analysis focuses on the transition to a specific state. It studies the distribution of the time to the transition or, more or less equivalently, the hazard of experiencing the transition. In the case of woman fertility, the interest is in the duration between successive parities; i.e., for example, the time to the first childbirth, the time from the 1st to the 2nd childbirth, from the 2nd to the 3rd childbirth, and so on.

2.2 Fertility histories as state sequences

We adopt in this paper an approach that differs from the descriptive and survival ones. We consider the fertility histories from a state sequence standpoint by deriving the successive yearly parity states—the number of children at the successive ages—from the time stamped childbirth events. States from age 0 to the age of the first childbirth are set to parity 0, and the parity is then increased by one at each new childbirth up to the fourth one. We distinguish thus between five states denoted P0, P1, P2, P3, and P4+ which constitute the alphabet. Each individual fertility history is characterized by an ordered list of yearly states from the above alphabet. Since we have state sequences, the position in the sequence informs about the elapsed time from the beginning of the sequence. It indicates also the age at that position.

For our analysis, we consider fertility histories from ages 15 to 50; i.e., sequences of length 36. The starting state indicates the parity at the 15th birthday. Two women in our sample had childbirth at 14 years, and, hence, have a fertility sequence starting in state P1. All other sequences start with state P0.

To make our data representation clear, we depict in Figure 1 the first four sequences of our data set in the SPS text form, which lists the distinct successive states with their duration, and graphically with an index plot which renders each sequence with horizontally stacked boxes colored according to the represented state. The latter plot clearly exhibits the timing and spacing of the successive childbirths. The first women stayed in state P0 (parity 0) during 21 years (from her 15th birthday to her 35th birthday), and then in state P1 during the rest of her fertility history; that is, during 15 years. Such a representation carries

```
    Sequence
[1] P0/21-P1/15
[2] P0/8-P1/1-P2/2-P3/15-P4+/10
[3] P0/11-P1/25
[4] P0/12-P1/6-P2/18
```

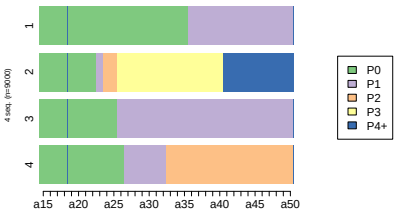

Fig. 1: The first four parity sequences of our data set

exhaustive and easily retrievable information about each fertility history: completed fertility, child spacing and age at each of the childbirths.

A peculiarity in this application of state sequence analysis to fertility histories is that the transitions between states are irreversible and can only occur in a predefined order. Therefore, the sequencing of the distinct states is not a concern, unlike, for instance, in the study of professional careers or family life courses.

3 Basis of state sequence analysis

Before turning to the search of representative sequences, we recall in this section some elements of state sequence analysis.

3.1 Aggregated view

A series of methodological tools can be used for analysing sets of state sequences. A first insight is provided by position-wise transversal characteristics (Gabadinho et al., 2011a). The sequence of transversal state distributions by cohort shown in Figure 2 depicts, for each cohort, the distribution of the cumulated parity at the successive birthdays. The results comply, for example, with the findings from the Swiss family fertility survey (FFS) that are reported in Gabadinho and Wanner (1999).

Since we have only complete trajectories, the proportion of women remaining in state P0 at each successive age is equivalent to the proportion of 'survivors' that would result from an event history analysis of the first childbirth. In addition, we observe, for the youngest cohorts in Figure 2, a reduction of the proportion of women experiencing more than three childbirths. This reduction is compensated mainly by an increase in the final proportion of parity 2 states.

The previous results could have been derived as well from repeated independent cross-sectional surveys, and, therefore, are not true longitudinal results. They give an aggregated overview, but do not render the individual characteristics—timing and spacing between successive childbirths—of the childbirth patterns. The sequence of transversal distribution is a kind of average representation, which hides any information on the structure and diversity of the individual patterns.

3.2 Longitudinal patterns

A true longitudinal approach has to account for the individual trajectories and their diversity. This requires considering each sequence as a whole unit of analysis. As a consequence, we have to deal with a high number of different possible patterns, even if only a small share of all theoretically possible sequences is actually observed. For our fertility histories, the predefined order in which transitions between states can occur limits the total number of possibilities. Nevertheless, given the 36 different ages at which the four possible transitions can occur, the the number of distinct patterns among the 1500 women of each cohort varies,

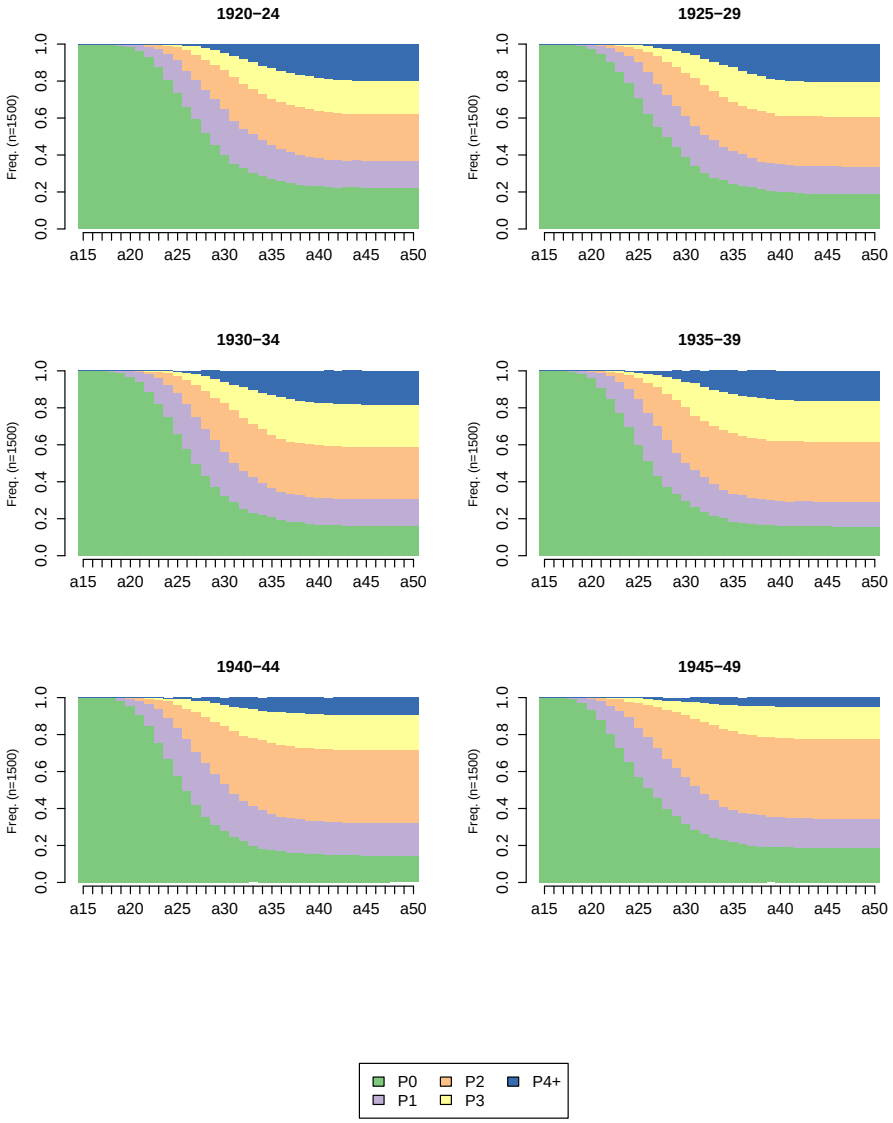

Fig. 2: State distribution plot.

as shown in Table 1 from 469 to 654. Overall, there are 1870 distinct patterns among the 9000 considered cases.

We can see the diversity of the patterns with an index plot of all individual sequences. Figure 3 plots the individual sequences for each of the cohort. To increase the readability of the plot, the displayed sequences are sorted by

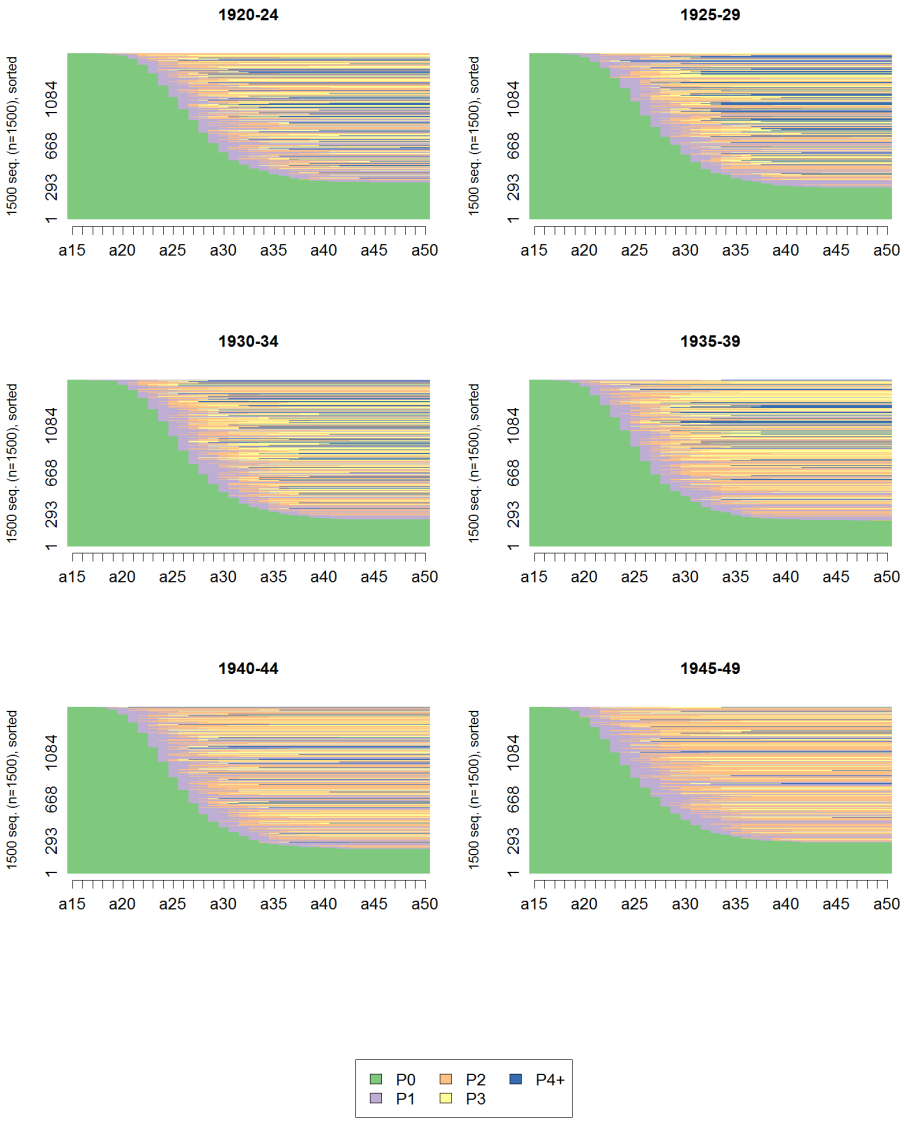

Fig. 3: Plot of individual sequences sorted by distance to the most frequent sequence.

their distance to the most frequent sequence. Despite this, each graphic remains burden. The plots convey so much information that it becomes almost impossible to characterize the typical patterns of each cohort through a simple visual inspection.

3.3 Pairwise dissimilarities between sequences

The start point of most advanced sequence analyses is the pairwise distance or dissimilarity between sequences. We already used such a distance to sort the sequences in Figure 3. From pairwise dissimilarities, we can, among others, measure the discrepancy—variance—of a set of sequences, cluster the sequences, run ANOVA-like analyses and grow regression trees for sequences (Studer et al., 2011). In the remaining of this chapter, we will need dissimilarity measures to find the medoid—the most central observed sequence—and define the neighborhood of a sequence.

Several dissimilarity measures or metrics have been proposed in the literature for state sequence data (see for example Abbott and Hrycak, 1990; Dijkstra and Taris, 1995; Elzinga, 2007; Lesnard, 2010). The most common measure in social sciences is optimal matching (OM) (Abbott and Forrest, 1986). OM is an edit distance that defines the distance $d(x,y)$ between two sequences x and y as the cost of transforming x into y by means of substitutions, insertions and deletions of states. The value of $d(x,y)$ depends indeed on the substitution and indel (insertion or deletion) costs. The indel cost is generally set as a constant while substitution costs often depend on the involved states.

For our analysis, we used a unit indel cost and a constant substitution cost of 2, in which case OM corresponds to the LCS distance, which can nicely be interpreted in terms of number of common attributes between the two sequences. The LCS distance between x and y is based on the length $A(x,y)$ of the longest common subsequence and is formally defined as

$$d(x,y) = A(x,x) + A(y,y) - 2A(x,y) \tag{1}$$

where $A(x,x)$ just is the length of x. The less common elements between two sequences, the greater the distance. The theoretic maximum distance is reached when the length of the longest common subsequence is 0, in which case the distance is twice the sequence length; i.e., $2 \cdot 36 = 72$ for our illustrative data.

The length of the longest common subsequence between the first two sequences in Figure 1 is 9, and the length of the sequences is 36. This yields a LCS distance of $36 + 36 - 2 \cdot 9 = 54$, which is quite large and confirms the great differences that we can observe between those first two sequences.

4 Typical sequences: Open issues

Now that we have recalled the basic concepts of sequence analysis, we can turn to the search of typical sequences. The aim of such *ideal types* or *ideal typical sequences* is to provide an easily understandable way of describing how individual trajectories typically look out. In our application, for example, we are interested to find, for each cohort, a sequence that would give an idea of the cumulated parities and the timing of the successive childbirths in the cohort. Before we propose our own solution, we begin by reviewing the approaches considered so far in the literature.

The question of finding an *ideal type* or *ideal typical sequence* is present, in the social science literature, in many articles that deal with sequence data, from the seminal work of Abbott and Forrest (1986) to recent applications (Aassve et al., 2007; Martin et al., 2008). An *ideal type* of a group of sequences is defined as a sequence that is a *typical version of* (Abbott and Hrycak, 1990), *a model for* the sequences in this group (Abbott, 1990), a sequence that best *represents* or *describes* the whole group (Aassve et al., 2007), or best summarises its *defining characteristics* (Martin et al., 2008). The ideal type helps the analyst to *separate the common pattern from its realizations* (Abbott and Forrest, 1986).

There are manifold approaches to derive the typical sequence. Though each approach assumes a somewhat different underlying meaning of 'typicality', they all follow the same goal: Find the sequence that best represents the whole set of sequences. We attempt to categorize these approaches and pinpoint their limitations.

Human versus data driven decision The identification of a typical sequence is most often just based on the inspection of the sequences by the researcher (Abbott, 1990; Martin et al., 2008). The tools used for inspection are rarely mentioned. We can easily imagine, however, that the inspection relies on graphical representations such as sequence index plots, state distribution plots or sequence frequency plots (Müller et al., 2008).

Figures 2 and 3 give two such graphic views of our data. Although they provide interesting general insights on the childbirth histories in each cohort, they do not help much for identifying a typical sequence. The state distribution plots in Figure 2 give an aggregated view that hides the individual patterns, while the sequence index plots in 3 are too burden to permit a clear identification of key regularities. Indeed, graphical inspection may be useful for small sets of sequences only, as is implicitly assumed by Abbott (1990, p. 165) who states: "Since the groups were small, [the ideal types] were derived by inspection".

Another obvious drawback of human inspection is the subjectivity of the solution; i.e., two domain experts may define different typical sequences. This is indeed problematic when reproducibility is a concern.

Few approaches propose automatic-data-driven selection procedures that do not require human decision and can reliably be reproduced. Such approaches consist in finding the sequence that optimizes some criteria, and the main point is, therefore, the definition of the criteria. In two of his papers (Abbott, 1990; Abbott and Hrycak, 1990), Abbott suggests to select "the sequence that minimizes some (possibly weighted) function of the distances to all other sequences", but does not use this approach in his own analyses. In a more recent work, Aassve et al. (2007) propose two alternatives to define the "ideal-typical" sequence of a cluster. Their first alternative is the *modal state sequence*; that is, a sequence built by taking at each position the modal state at that position. Their second and preferred alternative is the *medoid*, which is indeed Abbott's proposition. Robette (2010) also considers the *medoid* as a typical trajectory. The interest for the *medoid* comes from cluster analysis where it is commonly used as a short

cluster description (Kaufman and Rousseeuw, 2005). Interestingly, it can also be shown that the most central observed sequence is the nearest from the 'virtual' true center of the set (Studer et al., 2010).

Hypothetical versus observed sequences Human decision based solutions are not necessarily observed sequences. They may be artificial constructs reflecting hypothetical ideal types. This is also the case of some automatic-data-driven solutions, such as the sequence of modal states, for example. We can reasonably expect that humans will propose only plausible solutions. Automatic-data-driven procedures may end up, however, with non realistic solutions; e.g., a sequence where states 'parity 2' are followed by a state 'parity 1', which is impossible by definition. Aassve et al. (2007) mentioned this problem with regard to their *sequence of modal states* alternative: "The modal vector is not necessarily an observed sequence, implying that it may be inconsistent and therefore cannot be seen as an ideal-typical sequence". We could imagine to fit a statistical model which would generate the sequences under a series of consistency constraints. This is a hard task and no satisfactory solution does currently exist. Since, as argued by Aassve et al. (2007), it is essential that the typical sequence be a real one when describing results, we favor approaches that seek for the ideal types among the observed sequences.

Quality of the typical sequence As already mentioned, the aim of typical sequences is to represent at best the whole group of sequences. Therefore, we need a measure of the quality of representativeness to identify the best such representative. A first idea is to look at the distances between the chosen representative and all other sequences in the group. Aassve et al. (2007) compute, for example, the mean and maximum distance to the medoid, and a—mysterious and unexplained—proportion of the total dissimilarity accounted for by the medoid.

Mean and maximum distances are not totally satisfactory representativeness indicators since they do not account for the distribution of the sequences around their representative. Imagine, for instance, sequences organized in two clusters and an isolated sequence in-between at almost a same distance from each of the two clusters. This isolated sequence would be the most central and would therefore be chosen as the ideal type on the basis of the mean distance criterion, despite it is in fact an exception rather than a typical sequence. This example also raises the question of whether a single representative sequence is enough to represent the diversity of the patterns in the set. Clearly we would need at least two for our two cluster example.

5 A framework for finding a set of typical sequences

We have seen that the existing methods proposed in the literature for finding typical sequences face different limitations: subjectiveness for human-inspection-based solutions, possible inconsistency for artificial constructs, and possible low

representativeness for single sequence solutions. To overcome those disadvantages, we propose a data-driven method to find a subset of the observed patterns that is as small as possible and that can represent the whole set of sequences with a given level of accuracy. The method can be modulated by a series of possible representativeness criteria; therefore, is can be seen as a general framework.

5.1 Neighborhood, coverage and redundancy

The main idea behind our method is that, to be representative, a set of typical sequences must have at least a given percentage of the sequences in its neighborhood. The *neighborhood* is thus a key notion in our method. For a single sequence, it is the subspace within a given distance from the sequence and thus depends on the chosen distance measure. We call absolute and relative *coverage* of a pattern respectively the number and the percentage of sequences in its neighborhood. For a set of sequences, the neighborhood is the union of the neighborhoods of the sequences. The *coverage level* of a set of sequences is thus the percentage of cases that are within the neighbourhood of at least one of the pattern in the set.

Although the number of distinct patterns present in a set of state sequences is usually high, many of them are often just small variations of a same pattern and could be considered as realizations of a common pattern (Abbott and Hrycak, 1990). For example, the only difference between the two sequences below taken from our data set is that the first childbirth occurs one year later in the second sequence.

```
    Sequence
[1] (P0,15)-(P1,3)-(P2,18)
[2] (P0,16)-(P1,2)-(P2,18)
```

The two sequences obey a same general pattern. The second sequence is considered *redundant* since it is close to the first one, meaning that it lies in the neighborhood of the first pattern, or that it is covered by the first pattern.

5.2 A heuristic for finding a representative set

The idea of representative set comes from the biological sciences (Hobohm et al., 1992; Holm and Sander, 1998) where the aim is to get a reduced reference base of protein or DNA sequences in order to optimize queries in the sequence database. In this setting, the focus is on eliminating redundancy from the database. The representative set is expected to have "maximum coverage with minimum redundancy"; i.e., it must cover all the spectrum of distinct sequences present in the data, including outliers or rare sequences.

Our goal is similar regarding the elimination of redundancy. It differs, however, in that we consider representative sets with a user controlled coverage level; i.e., we do not require maximal coverage. We thus define a representative set as a subset of non redundant 'typical' sequences that largely, though not necessarily exhaustively, covers the spectrum of observed sequences.

Due to the possibly huge number of distinct patterns, searching exhaustively for the best set of representative sequences is not feasible. We therefore propose a heuristic approach. In a first stage, we sort the distinct observed patterns, and we then select the representative by removing successively from the list patterns redundant with already selected ones. Although, this process does not necessarily end up with the smallest set of representatives for a given coverage, it provides always almost optimal solutions. Basically, the heuristic requires the user to specify a representativeness criterion for the first stage and a similarity threshold for evaluating redundancy in the second one. The heuristic can be summarized as follows:

1. Compute a representativeness score for each distinct pattern in the data set;
2. Sort all distinct patterns according to their score, in ascending order if the score increases with representativeness or in descending order otherwise;
3. Starting with the most representative pattern, keep iteratively from the list each sequence for which the dissimilarity with any already retained sequence is greater than a given threshold. Stop, when the expected overall coverage is attained.

By construction, the final set of representative sequences will not contain any pair of sequences closer from each other than the predefined threshold.

5.3 Controlling the Size/Coverage Trade-off

The heuristic can be fine tuned by means of several parameters and through the choice of the initial representativeness sorting criterion.

Representativeness criteria. The solution provided by the heuristic depends on the initial sort of the sequences. Different representativeness criteria can be considered, each having its own interest. We shortly comment hereafter three possibilities: the *neighbourhood density*, the *frequency* and the *centrality*. The latter two use the retained dissimilarity measure.

Frequency. It is the number of occurrences of the pattern. The more frequent a sequence, the more representative it is supposed to be.

Neighborhood density. It is the coverage of each single sequence for a specific sorting neighborhood radius. The higher the density, the higher the representativeness. The neighborhood density generalizes the frequency, which is the density for a zero neighborhood radius.

Centrality. It is the sum of the distances to all sequences in the set; that is, the criteria used for finding the most central sequence—the medoid (Kaufman and Rousseeuw, 2005). The smallest the mean distance, the most representative the sequence.

Control parameters. Parameters that can be used to control the tradeoff between the number of representatives and the coverage are: the redundancy threshold, the neighborhood radius for the overall coverage and the minimal overall coverage level. Increasing the latter for a given neighborhood radius will increases the size of the representative set, while increasing the radius for a fixed minimal overall coverage will reduce the size of the set. Instead of setting the minimal wanted overall coverage level, we can fix the wanted number of representatives.

The radius for the overall coverage and the redundancy threshold will generally be set as the same value, but different values could be used as well. When the neighborhood density is used for the initial sort, a specific radius can also be set for that. A convenient way to specify those radii is to set them as a proportion of the maximal theoretical distance between two sequences.

Let us illustrate with our 36 year long sequences, for which the theoretical maximum LCS distance is $2 \cdot 36 = 72$. Setting the redundancy threshold as 10% of this theoretical maximum yields a threshold of 7.2. Two sequences are then considered redundant when the length of their longest common subsequence exceeds $(72 - 7.2)/2 = 32.4$; i.e, is 33 or greater. Since the sequence length is 36, we have redundancy as long as there are only three or less non common states.

6 Quality of the representative set

Depending on the parametrization of the method, different representative sets will result. It is therefore important to evaluate the typicality or representativeness of the obtained set of patterns. We propose hereafter a few such quality measures.

Let $r_1, \ldots, r_k$ be the k representative patterns. We assign each sequence x to its closest representative and denote by R_i the set of indexes of the sequences assigned to r_i, and by $a_i = |R_i|$ the number of those assigned sequences. When a sequence is equally distant from two or more representatives, we assign it to the pattern with the highest representativeness score. Letting n be the total number of sequences, we have $n = \sum_{i=1}^{k} a_i$. We can now define the following measures.

Mean Distance. A first quality measure of the representative r_i is the mean distance between r_i and its a_i assigned sequences

$$MD_i = \frac{SD_i}{a_i}$$

where $SD_i = \sum_{j \in R_i} d(x_j, r_i)$ is the sum of distances to the assigned sequences. The closer r_i is from its assigned sequences, the better it is.

Overall coverage and contribution to the coverage. A second quality indicator of a representative r_i is the number of sequences among those assigned

to r_i that are in its neighbourhood; i.e., within a distance δ

$$b_i = \sum_{j \in R_i} \Big(d(x_j, r_i) < \delta \Big) \ .$$

where δ is the neighborhood radius for the overall coverage. The quantity b_i is the contribution of the ith representative to the overall absolute coverage $b = \sum_i^k b_i$. In proportion of the number n of sequences, the overall coverage is b/n.

Distance gain. A third way of measuring the quality of a representative pattern r_i is by comparing the sum SD_i of distances from r_i to its assigned sequences with the sum $DC_i = \sum_{j \in R_i} d(x_j, c)$ of the distances from the true center c of the whole set to these same a_i assigned sequences. The idea is to measure by how much the representative r_i is closer to those sequences than the center c. We define thus the quality measure Q_i of r_i as

$$Q_i = \frac{DC_i - SD_i}{DC_i}$$

which is the reduction rate of the sum of distances. Q_i may be negative in some circumstances, when there is a single representative for example, meaning that the sum of the a_i distances to the representative r_i is higher than the sum of the a_i distances to the true center c of the set.

In the same vein as Q_i, we assess the overall quality of the representative set with

$$Q = \frac{\sum_i^k DC_i - \sum_i^k SD_i}{\sum_i^k DC_i} = \sum_{i=1}^{k} \frac{DC_i}{\sum_{j=1}^k DC_j} Q_i \ .$$

Representing all sequences by a sequence located exactly at the center of the set yields $Q = 0$.

Discrepancy. A last measure of interest of a representative r_i is the discrepancy within the set R_i of its assigned sequences. We measure this discrepancy with the sum $SC_i = \sum_{j \in R_i} d(x_j, c_i)$ of distances from the true center c_i to the a_i sequences assigned to r_i, or with the mean of those distances $V_i = SC_i/a_i$ (see Studer et al., 2010, for details on how to compute the discrepancy).

7 Application

We now illustrate the method with our data set. We first try to find a single typical sequence by cohort and then look for a small set of representatives for each cohort that covers 50% of its trajectories. Beside the demographic knowledge about the evolution of the fertility patterns across cohorts, the application also highlights the effect of the tuning parameters.

The analysis is based on the LCS pairwise distances (see Section 3.3) between fertility trajectories. We used, for the overall coverage, a neighborhood radius of

10% of the maximal possible distance. When applicable, we retained the same value as redundancy threshold and as radius for measuring the density around a pattern.

7.1 Looking for a single representative

We begin by studying the solutions obtained with the different sorting criteria when we fix the size of the representative set to 1; i.e., when we want only one representative pattern per cohort. The sorting criterion is in that case of primary importance, since the single solution is, with our heuristic, just the first pattern in the sorted pattern list.

Figure 5 shows, for each cohort, the typical sequence identified with each of the three sorting criteria and Table 3 reports coverage and centrality statistics for those single solutions. Childlessness, i.e., staying in state P0 from age 15 to 50, is the most frequent trajectory in all cohorts, while the most central pattern corresponds to two childbirths and a similar intergenesic interval of about 4 years in all groups. Only the timing differs across cohorts, the first childbirth occurring about 5 to 7 years earlier (28-29 versus 35-36)[2] in the youngest cohorts than in the oldest ones. The pattern with highest density, on its side, is equal or very close to the most frequent pattern for the first 4 cohorts, while it resembles to the most central one for the youngest two cohorts.

Frequency and density give the better results in terms of coverage, whereas centrality generates the best solution in terms of overall distance to the whole set of sequences; i.e., solutions with lower values for the MD statistic. Since we have only one representative, the Q statistic which reflects by how much the representative is closer from the sequences than the true virtual center can only be negative. It also is closer to zero for the most central sequence.

Figure 4 gives a scatterplot representation of the sequences in the oldest and youngest cohorts. The—principal—coordinates of the sequence-points were obtained through a multidimensional scaling analysis of the pairwise dissimilarity matrix. The idea is to represent the sequences in a two-dimensional real space, such that the Euclidean distances between pairs of points in the real space conform as much as possible to the pairwise dissimilarities between the sequences. The points are colored according to the frequency of the corresponding pattern. The darker the point, the more frequent it is. The plot also shows where the three representative patterns are located and how they position themselves with respect to the other sequences. We see that the most frequent pattern is clearly off-centred, while the most central one tends to be located on the border of a cluster. The pattern with the highest density is centered in a cluster in the youngest cohort, whereas it is off-centered in the oldest cohort.

Let us look in more details at each solution.

[2] Statuses are those at a given anniversary, and thus changing from status P0 to P1 between ages 28 and 29 means that the childbirth occurred at age 28.

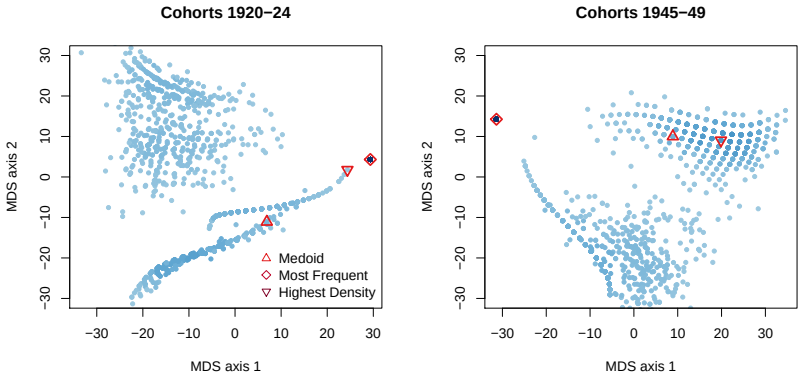

Fig. 4: Most central, frequent and dense patterns projected on the first two principal coordinates. Cohorts 1920-24 and 1945-49

Most frequent sequence The most frequent pattern, childlessness, is the same in all groups, which is not surprising since it is the sole possible trajectory for women staying childless. Depending on the cohort, the relative frequency of this pattern ranges from 14.6% in cohort 1940-44 to 22.1% in the cohort 1920-24. Interestingly, although staying childless is the most common trajectory in all cohorts, childlessness is the most frequent completed fertility—end state—in none of the cohorts. This is because there is no variability possible among women staying childless, whereas there are many different trajectories leading to any other final state. There are, for instance, 263 distinct patterns that end with P2; i.e., with a total of two childbirths.

The coverage of the most frequent sequence in each cohort does not exceed its frequency by more than 0.1%. It is, for example 22.2% versus 22.1% for the first cohort. Covered sequences include childlessness trajectories and trajectories with childbirths after 47 years old; i.e., sequences with a childlessness spell of at least length 33 (as computed in Section 5.3).

Most central pattern. We have seen that the medoid is a two childbirths pattern for every cohort. This looks consistent with the highest proportions observed in Table 1 for the P2 completed parity in each of the cohort. However, the coverage of the medoid is poor compared to the most frequent sequence, especially for the four oldest cohorts. Only 2 to 9% of the sequences are covered; i.e., distant from less than 10% of the maximum theoretical distance from the associated medoid, as compared to the 22 to 16% of sequences covered by the most frequent sequence. The low coverage obtained for the medoid means that it is located in a low density area (in the space defined by the distance matrix), and that it is, therefore, a rather atypical pattern.

Table 3: The most typical sequence for 3 different criteria.

	seq.	*nb*	(%)	*MD*	*V*	*Q*
1920-24						
Frequency	(P0,36)	333	22.2	36.2	18.6	-94.7
Density	(P0,33)-(P1,3)	337	22.5	33.8	18.6	-82.1
Centrality	(P0,22)-(P1,4)-(P2,10)	30	2.0	30.7	18.6	-65.6
1925-29						
Frequency	(P0,36)	287	19.1	37.7	18.8	-101.1
Density	(P0,36)	287	19.1	37.7	18.8	-101.1
Centrality	(P0,20)-(P1,4)-(P2,12)	59	3.9	30.9	18.8	-64.5
1930-34						
Frequency	(P0,36)	242	16.1	40.1	19.0	-110.9
Density	(P0,36)	242	16.1	40.1	19.0	-110.9
Centrality	(P0,17)-(P1,5)-(P2,14)	77	5.1	31.5	19.0	-65.7
1935-39						
Frequency	(P0,36)	236	15.7	41.5	19.3	-115.4
Density	(P0,33)-(P1,3)	240	16.0	38.4	19.3	-99.5
Centrality	(P0,16)-(P1,4)-(P2,16)	119	7.9	31.3	19.3	-62.5
1940-44						
Frequency	(P0,36)	220	14.7	42.3	19.0	-122.1
Density	(P0,11)-(P1,3)-(P2,22)	293	19.5	31.0	19.0	-62.9
Centrality	(P0,14)-(P1,4)-(P2,18)	196	13.1	29.7	19.0	-56.1
1945-49						
Frequency	(P0,36)	279	18.6	40.6	18.7	-117.3
Density	(P0,10)-(P1,3)-(P2,23)	309	20.6	30.7	18.7	-64.3
Centrality	(P0,15)-(P1,4)-(P2,17)	173	11.5	28.7	18.7	-53.4

In cohort 1920-24, for example, the median age at the first childbirth is 27 years for women ending their reproductive life with a completed fertility of 2 children (see Table 2), and the third quartile is situated at 30 years. This is much earlier than the first childbirth at 36 years that we have in the medoid pattern. The latter corresponds, therefore, to an unusually late 2 children fertility schedule. The same observation can be done for the other cohorts.

Highest neighbourhood density. By definition, since we use the same radius for the density and the coverage, the pattern with highest density also has the highest coverage. We have seen that the most densest pattern is similar to the most frequent one in cohorts born before 1940, in which case its coverage is equal or slightly greater to the most frequent pattern. In the youngest two cohorts, the

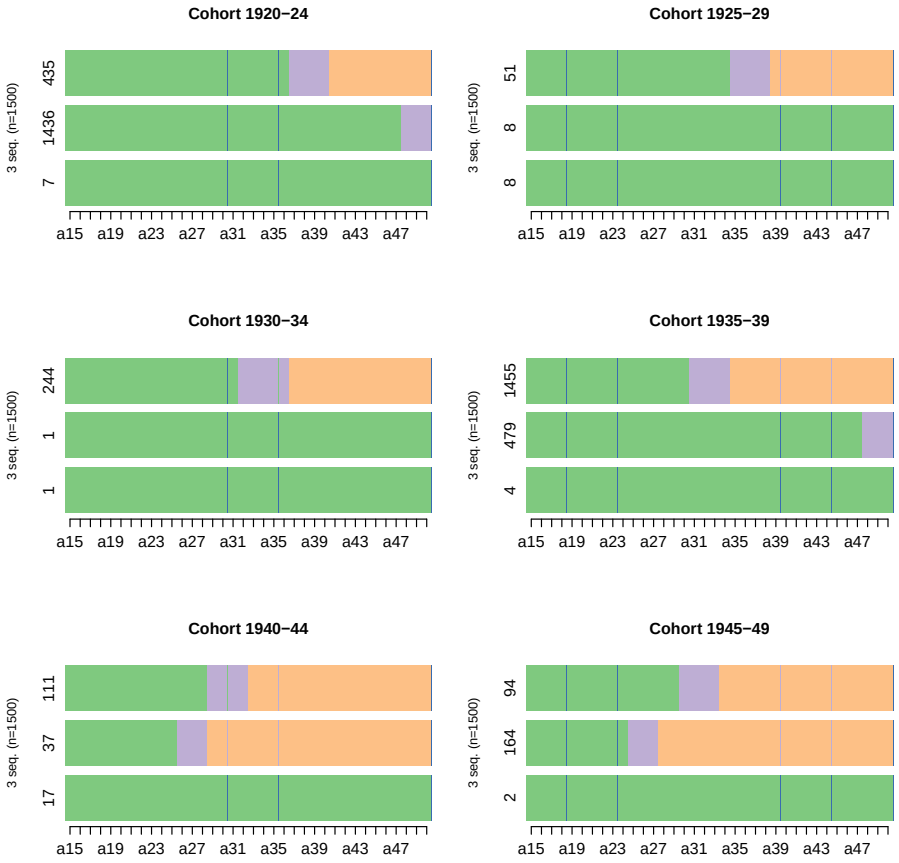

Fig. 5: Most frequent sequence, sequence with highest neighbourhood density and most central sequence (medoid), bottom up, for each cohort.

pattern with highest density is a two childbirths pattern like the most central one. Unlike the medoid, the pattern with highest density is, in the youngest cohorts, typical, in that the ages at the childbirths are close to the median age observed for women ending their reproductive life with two children. The pattern with highest density also has significantly better coverage than the two other solutions in those two cohorts.

We can conclude that none of the single-representative solutions is really satisfactory. The most frequent pattern—childlessness—does not reflect that two childbirths are more common than none. This is because it does not account for the similarity between patterns with childbirths at slightly different ages. The

most central pattern exhibits atypical late childbirths, which is a consequence of the compromise between no childbirths at all and multiple childbirths. The pattern with highest neighborhood density shares some of the drawbacks of the frequency, especially when it is based on a small radius. On our data, for example, the pattern with highest density is close to the most frequent one in four out of the six considered groups. It is more consistent for the two youngest cohorts.

7.2 Looking for a set of representatives

Since single-pattern solutions can not ensure a satisfactory coverage, it is natural to look for a set of representative patterns rather than for a single one. We ran therefore our heuristic by setting a 50% minimal coverage. The obtained sets by cohort are displayed in Figure 6. The width of the horizontal bars in the representative plot is proportional to the number of sequences assigned to the corresponding pattern. Their order reflects the position of the pattern in the initial sorted list.

Interpretation. The representative patterns clearly exhibit the shift from a mix of patterns with four, two, one or no childbirths in the oldest cohorts to the predominance of two childbirth patterns in the youngest cohorts. While patterns with four childbirths are typical of women born before 1935, and patterns with three childbirths are common for women born between 1930 and 1939, such trajectories are no longer representative of women born after 1940. We can see also that the representativeness of trajectories with no childbirths remains more or less the same across cohorts.

Although, similar findings can be we drawn from Figures 2 and 3, they are much clearer here. We get a crisper image of what typical fertility histories are. For example, the fact that trajectories with more than two childbirths disappear from the set of characteristic patterns of the younger cohorts is very instructive. Furthermore, when compared with the distribution plots in Figure 2, the representative plots render the diversity of the common patterns. They also are easier to read than the burden i-plots in Figure 3. Interestingly, we observe, for example, a diversity in the birth timing among trajectories with two childbirths, which seems to persist across cohorts.

The obtained representative sets provide a much more realistic picture than the single solutions addresses in Section 7.1. The results are consistent with the distribution of the completed parities in Table 1 and the timing statistics in Table 2. For example, the successive drift from patterns with four childbirths to patterns with three and then two childbirths conforms to the regular decrease in the completed P4+ parity, and the reversed-V evolution of the percentages in the completed P3 parity that can be observed in Table 1. The typical childbirth timing that we observe in Figure 6 also conforms to the median ages in Table 2. For example, if we look at the four-child pattern in the first cohort, we see that the first two births occur respectively at 26 years and 27 years, while the median ages reported in Table 2 for completed parity R4+ are 25 and 27 years.

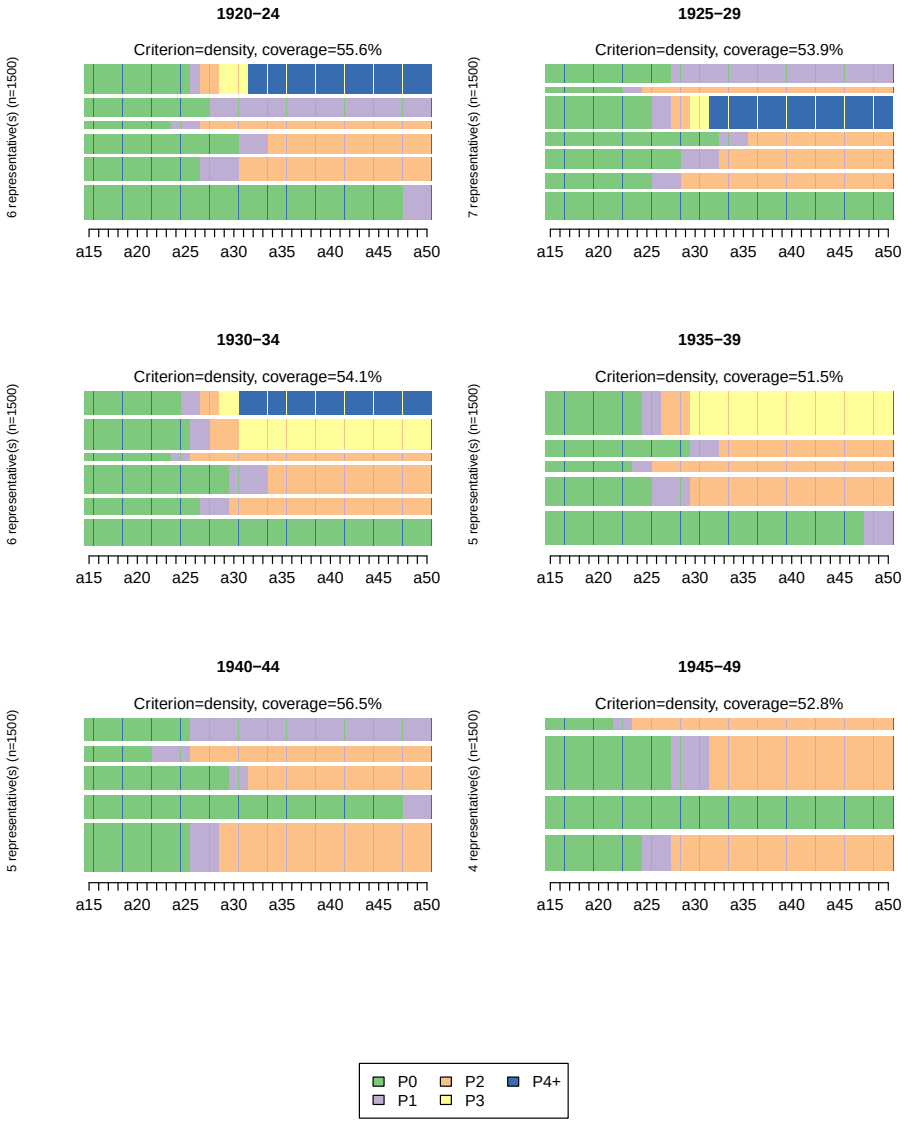

Fig. 6: Representative sequences by cohort for a 50% minimal coverage, density criterion.

Changing the control parameters. As already mentioned, the search algorithm can be controlled by a series of parameters which affect the size of the resulting set of patterns and its representativeness. To get an idea of the effect of those parameters, we display, in Table 4, statistics about the representative sets

Table 4: Comparing criteria with coverage thresholds of 25%, 50%, 75% and 100%.

	k	(%)	*MD*	*Q*	*k*	(%)	*MD*	*Q*	*k*	(%)	*MD*	*Q*	*k*	(%)	*MD*	*Q*
1920-24																
Frequency	2	29.8	28.2	-52.1	6	50.9	16.8	9.3	13	75.4	4.7	74.5	89	100.0	2.4	87.0
Density	2	33.7	23.7	-27.9	6	55.6	12.1	34.6	14	75.1	6.2	66.6	94	100.0	3.6	80.6
Centrality	32	28.9	10.7	42.2	50	65.1	6.0	67.7	58	75.5	4.8	74.0	102	100.0	2.2	88.0
1925-29																
Frequency	3	28.7	29.3	-55.9	8	50.0	17.3	7.7	17	76.5	4.6	75.3	89	100.0	2.4	87.3
Density	2	29.9	24.7	-31.6	7	53.9	11.6	38.0	15	75.1	4.9	74.1	90	100.0	2.4	87.2
Centrality	23	25.1	13.7	26.9	44	50.3	8.7	53.4	54	76.5	4.9	74.0	97	100.0	2.4	87.3
1930-34																
Frequency	2	27.3	26.2	-37.8	8	51.7	18.3	3.7	15	78.0	4.7	75.2	88	100.0	2.5	86.6
Density	2	29.8	25.2	-32.6	6	54.1	10.9	42.8	16	76.3	5.1	73.3	93	100.0	2.5	86.8
Centrality	18	26.9	15.8	16.6	44	50.1	9.5	50.3	67	84.7	4.0	79.2	94	100.0	2.5	87.1
1935-39																
Frequency	3	37.0	20.0	-3.8	8	52.9	18.0	6.7	14	75.3	5.0	74.1	83	100.0	2.5	86.8
Density	2	31.7	24.7	-28.4	5	51.5	15.9	17.6	14	75.8	6.0	69.0	85	100.0	3.5	82.0
Centrality	15	25.9	17.9	6.9	46	50.9	8.1	57.9	71	87.0	3.8	80.2	96	100.0	2.5	87.1
1940-44																
Frequency	3	27.0	32.0	-67.9	5	53.1	15.8	16.8	13	77.5	7.2	62.0	75	100.0	2.5	86.9
Density	2	34.3	23.3	-22.3	5	56.5	16.0	15.7	14	77.3	8.0	58.0	78	100.0	3.3	82.5
Centrality	4	26.8	25.6	-34.3	36	54.5	9.1	52.0	59	75.0	5.1	73.1	86	100.0	2.4	87.3
1945-49																
Frequency	2	25.7	31.8	-70.0	4	56.5	13.2	29.3	13	78.1	5.9	68.2	77	100.0	2.3	87.8
Density	2	39.2	20.1	-7.6	4	52.8	17.9	4.4	11	76.8	6.1	67.4	76	100.0	2.3	87.9
Centrality	4	26.7	24.1	-28.9	31	52.0	9.3	50.0	49	81.5	4.8	74.1	88	100.0	2.2	88.1

obtained with each of the three considered initial sorting criteria and minimal wanted coverage of 25%, 50%, 75% and 100%. Radii for redundancy, neighborhood and coverage are all set as 10% of the maximal distance.

A first salient fact is that the centrality criterion (selecting the sequences according to their distance to the center of the set) yields a very low coverage/size ratio; i.e., a low coverage for the obtained number of representatives. A second salient fact is that the neighbourhood density criterion yields consistently the best coverage/size ratio for a wanted coverage up to 50%, while frequency gives sometimes slightly better results when the wanted coverage is high. For our data, where the number of different possible patterns is limited by the strict order in which states can occur, we have several patterns with a relatively high frequency. This explains why we get very similar results with the frequency

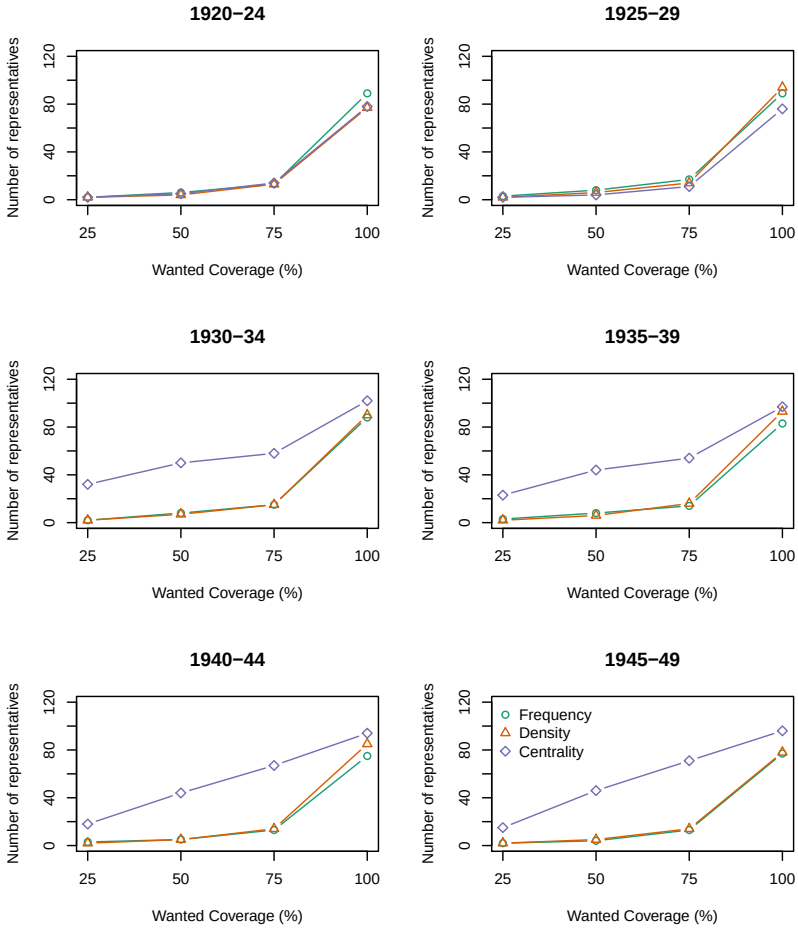

Fig. 7: Number of representative sequences versus required minimal coverage, by cohort and sorting criterion.

and density sorting criteria. Experiences with other sequence datasets exhibited a much clearer superiority of the density criterion over frequency (Gabadinho et al., 2011b).

Figure 7 illustrates the trade-off between the wanted coverage and the number of representatives. We also observe in that figure that the differences in size of the representative sets obtained with the three criteria diminish when we want a 100% coverage. The reason for this phenomenon is that there are fewer possibilities to achieve a complete coverage than a 50% coverage for example. The obtained representative set is, therefore, less dependent on the order in which we scan the observed sequences.

8 Conclusion

We addressed the question of identifying typical sequences that can characterize the most common features of observed life trajectories. We have seen that attempting to find a single typical sequence by looking, for instance, for the most central one or the sequence made of the position-wise modal states, is most often unsatisfactory because it can result in a pattern that either is not characteristic at all of the set or even is inconsistent. We proposed, therefore, to look for a set of representatives rather than for a single solution and developed a flexible method to find an as small as possible set of representatives covering a given percentage of the sequences. The method is based on the concept of neighborhood, the coverage being the number of cases in a given neighborhood of the representatives. We also propose a representative plot to render the selected typical patterns.

We applied the method on childbirth histories of Swiss women sampled from the 2000 Swiss federal census. The sets of representatives extracted for each of six birth cohorts are very appealing. They provide an easily interpretable picture of the most common patterns in each cohort and permit to easily understand the evolution in those patterns across cohorts. The representative plot enriches the average view provided by the—commonly used—plot of the successive transversal state distributions (d-plot), and renders somehow the diversity of the sequences than can be observed in the—often burden—plot of all individual sequences (i-plot). Furthermore, we have seen that the representative plot conveys useful information on typical timings. The results exhibit clearly a drift from cohorts with a strong presence of patterns with four or more childbirths to cohorts dominated by patterns with two childbirths, but also shows that there is an important birth-timing diversity of the two-childbirth patterns. Surprisingly, the same typical timing can be observed in all cohorts.

The aim of the application is mainly to illustrate the scope of the method. We based our analysis on the LCS dissimilarity measure, which is derived from the length of the longest common subsequence. This measure does not account for the closeness of two successive states, meaning that parities 0 and 1 are considered as different as parities 0 and 4. It could then be of interest to repeat the analysis with a more appropriate distance. Nevertheless, the findings clearly make sense.

A possible drawback of our method is that it requires some fine tuning to get a satisfactory representative set. Indeed, the set of representatives should not be too large to remain comprehensible, but should at the same time cover a high number of sequences. We have, therefore, to manage this trade-off and this may require some trials. We have proposed a series of quality measures to help in that process. Those measures play for our method a role similar to the criteria (Milligan and Cooper, 1985; Kaufman and Rousseeuw, 2005) used for determining the number of groups in cluster analysis. However, our approach aims to achieve good coverage and representativeness, while the goal of clustering is good partition. Our measures concern, therefore, coverage and centrality rather than the cohesion and separation of groups.

Finally, let us mention that the proposed tools are made available in our TraMineR R-package for categorical sequence analysis (Gabadinho et al., 2011a).

Bibliography

Aassve, A., F. Billari, and R. Piccarreta (2007). Strings of adulthood: A sequence analysis of young British women's work-family trajectories. *European Journal of Population 23*(3), 369–388.

Abbott, A. (1990). Conception of time and events in social science methods: Causal and narrative approaches. *Historical Methods 23*(4), 140–150.

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

Abbott, A. and A. Hrycak (1990). Measuring resemblance in sequence data: An optimal matching analaysis of musician's carrers. *American Journal of Sociolgy 96*(1), 144–185.

Billari, F. C. (2005). Life course analysis: Two (complementary) cultures? Some reflections with examples from the analysis of transition to adulthood. In R. Levy, P. Ghisletta, J.-M. Le Goff, D. Spini, and E. Widmer (Eds.), *Towards an Interdisciplinary Perspective on the Life Course*, Advances in Life Course Research, Vol. 10, pp. 267–288. Amsterdam: Elsevier.

Dijkstra, W. and T. Taris (1995). Measuring the agreement between sequences. *Sociological Methods and Research 24*(2), 214–231.

Elzinga, C. H. (2007). Sequence analysis: metric representations of categorical time series. Technical report, Department of Social Science Research Methods - Vrije Universiteit Amsterdam, The Netherlands.

Gabadinho, A., G. Ritschard, N. S. Müller, and M. Studer (2011a). 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 (2011b). 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.

Gabadinho, A. and P. Wanner (1999). *Fertility and family surveys in countries of the ECE region: standard country report, Switzerland.* New York.

Hobohm, U., M. Scharf, R. Schneider, and C. Sander (1992). Selection of representative protein data sets. *Protein Sci 1*(3), 409–417.

Holm, L. and C. Sander (1998). Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics 14*(5), 423–429.

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

Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods and Research 38*, 389–419.

Martin, P., I. Schoon, and A. Ross (2008). Beyond transitions: Applying optimal matching analysis to life course research. *International Journal of Social Research Methodology 11*(3), 179 – 199.

Milligan, G. W. and M. C. Cooper (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika 50*(2), 159–179.

Müller, N. S., A. Gabadinho, G. Ritschard, and M. Studer (2008). Extracting knowledge from life courses: Clustering and visualization. In I.-Y. Song, J. Eder, and T. M. Nguyen (Eds.), *Data Warehousing and Knowledge Discovery, 10th International Conference, DAWAK 2008, Turin, Italy, September 2-5*, Volume LNCS 5182 of *Lectures Notes in Computer Science*, pp. 176–185. Berlin Heidelberg: Springer.

Robette, N. (2010). The diversity of pathways to adulthood in france: Evidence from a holistic approach. *Advances in Life Course Research 15*(2-3), 89–96.

Schoumaker, B. and S. R. Hayford (2004). A person-period approach to analysing birth histories. *Population (English Edition) 59*(5), pp. 689–701.

Studer, M., G. Ritschard, A. Gabadinho, and N. S. Müller (2010). Discrepancy analysis of complex objects using dissimilarities. In F. Guillet, G. Ritschard, D. A. Zighed, and H. Briand (Eds.), *Advances in Knowledge Discovery and Management*, Volume 292 of *Studies in Computational Intelligence*, pp. 3–19. Berlin: Springer.

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

Tuma, N. B. and J. Huinink (1990). Postwar fertility patterns in the federal republic of germany. In K. U. Mayer and N. B. Tuma (Eds.), *Event History Analysis in Life Course Research*. The University of Wisconsin Press.

Wiggins, R. D., C. Erzberger, M. Hyde, P. Higgs, and D. Blane (2007). Optimal matching analysis using ideal types to describe the lifecourse: An illustration of how histories of work, partnerships and housing relate to quality of life in early old age. *International Journal of Social Research Methodology 10*(4), 259 – 278.

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