Describing Life Courses. An Illustration Based on NLSY Data

Götz Rohwer
Heike Trappe

February 1997

Paper prepared for the POLIS project conference at the European University Institute, Florence, February 28 – March 1, 1997.

Address of authors: Max-Planck-Institut für Bildungsforschung, Lentzeallee 94, 14195 Berlin.

Introduction

The central research questions of the POLIS project are how life courses have changed across successive birth cohorts, and whether part of these changes can be attributed to changes in institutional frameworks. Whatever the final approach to these questions will be, as a first step one needs suitable tools for describing life courses. Surprisingly enough, such tools are not well developed yet. Most sociological research in the event-history framework has its focus on specific types of events and individual transitions, not on the finally resulting life courses and trajectories.

From a methodological point of view, the dominating focus on transitions is quite understandable. Individual life courses are *the result* of going sequentially through time; metaphorically spoken, the tracks that are left behind when people grow older. Consequently, trying to explain the development of life courses one needs to investigate the transitions which might occur during life courses.

On the other hand, this approach easily runs into the danger of loosing sight of the final result, the life courses as resulting from a continuous stream of currently happening events, and providing a continuously changing identity to its individual „bearers". Given their sequential development it is, of course, questionable whether whole life courses can be reasonably taken as „dependent variables" (suggested, for instance, by Abbott 1995, p. 105). However, given that there is already a broad sociological literature taking life courses as the subject of its discourse, it should be an important task to develop empirical and statistical methods that not only allow to investigate individual transitions but can contribute to analytically useful descriptions of whole life courses. This seems to be particularly important in order to establish an empirical basis for the POLIS project.

In this paper we try to contribute to a discussion of methods for describing life courses which might be useful for the POLIS project.

- In section 1 we propose a simple data structure for representing life courses as sequences of states.
- In section 2 we illustrate this proposal by using life history data from the National Longitudinal Study of Youth (NLSY).
- It follows in section 3 a short discussion of cross-sectional distributions for describing the evolution of life courses. It is shown that this approach, although often used, can easily be misleading.
- A somewhat broader view, focusing on the question of how to establish longitudinal classifications, is taken in section 4. We investigate two methods for assessing the stability of group membership over time.
- We then begin with a discussion of describing life courses on an individual level. One simple approach, focusing on the occurrence of events, is treated in section 5. We discuss how this approach can also be used for repeatable and complex events (e.g., transition from education to work).
- Section 6 then raises the question whether we can hope to find typical careers and begins a discussion of some problems connected with this idea. The discussion is continued in Section 7 where we use a simple clustering procedure to illustrate the difficulties in searching for typical careers.

1 Representing Life Courses

We begin with shortly describing a formal framework for representing life courses. The basic idea is to represent individual life courses as sequences of states. We assume a basic unit of time, say weeks or months, and a state space providing a set of different states such that each individual is in exactly one of these states during each of the basic time periods.[1]

This immediately leads to a simple formal representation of life courses. Let $\mathcal{Y}$ denote the state space, and $t = 1, 2, 3, \ldots$ be an index for the sequence of time units. The sample of individuals will be indexed by $i = 1, \ldots, N$, and y_{it} will be used to denote the state of individual i during the time unit t. We then get representations of the individual life courses by the sequences

$$y_i = (y_{i1}, y_{i2}, y_{i3}, \ldots)$$

What can be represented depends on the definition of the state space and the time axis. However, the framework is quite flexible. There are no theoretical limits in making the state space more and more differentiated. Alternatively, we can define two or more separate state spaces. For instance, a state space $\mathcal{Y}$ can be used for education and labor market activities, and a state space $\mathcal{Z}$ can be used to record partnerships

[1] This approach is somewhat different from the conventional view of event-history data as sequences of episodes, or spells. However, when based on the same time units as used for measuring the dates of events, both approaches provide identical information. This becomes different only if the definition of sequence data uses more aggregated time units. As an example, see Gershuny (1993, p. 141) who has used sequences generated by observing life courses once per year. This might be sufficient for investigating occupational mobility, but not, for instance, for investigating transitions into and out of unemployment.

and family affairs. Representation of individual life courses is then by two-dimensional sequences

$$(y_i, z_i) = (y_{i1}, z_{i1}, y_{i2}, z_{i2}, y_{i3}, z_{i3}, \ldots)$$

While it would be possible to combine both dimensions into a single state space (the cartesian product of $\mathcal{Y}$ and $\mathcal{Z}$), keeping distinct state spaces provides an opportunity for investigating temporal relationships between events in the marginal processes.[2]

In our view, this formal framework for representing life courses is particularly well suited for the POLIS project.

- It provides a general and flexible framework for representing most aspects of life courses which seem important for the research questions of the POLIS project.
- While the representation of life courses as sequences of states directly draws attention to whole life courses, the same framework can be used to investigate specific transitions with conventional methods of event history analysis.
- The representation of life courses as sequences of states not only provides a unifying data structure for cross-country comparison of life courses, but also creates a direct link to the theoretical questions of the POLIS project regarding the existence of „typical“ life courses and their changing features across cohorts.

In the following sections we use this sequence data structure as a framework for discussing methods for describing life courses. All calculations and plots have been done with the computer program TDA that supports this data structure (see Rohwer, 1996).

2 Illustration with NLSY Data

To illustrate the sequence data structure introduced in the previous section, we use data from the National Longitudinal Study of Youth (NLSY). This is a nationally representative sample for the birth cohorts between 1957 and 1964 in the United States, consisting of 12,686 young women and men who were first surveyed in 1979. Interviews with NLSY respondents have been conducted yearly since 1979 (panel study). The NLSY sampling design enables researchers to study in detail the longitudinal experiences of not only this particular age group of young Americans but to analyze the disparate life course experiences of such groups as women, hispanics, blacks, and the economically disadvantaged (see Center for Human Resource Research, 1994).

To illustrate the mainly methodological discussion in this paper, we focus on respondents born in 1964, that is, we use only a single birth cohort. As a consequence, we do not explicitly discuss questions of cohort comparison. However, many of the methods discussed below can be used for this purpose.

In our data set, the last interview is in 1990. There are 1106 respondents born in 1964, most of them with an interview in 1990.[3] For these respondents we create sequence data based on a monthly time axis beginning for each respondent with the month of his or her 15th birthday: $t = 0, 1, 2, \ldots$ ($t = 0$ is month of 15th birthday, $t = 1$ is the following month, and so on until the month of the last interview). The time axis runs until $t = 142$, that is, the longest possible observation period is 143 months (about 12 years, = age 27).

For our illustratory purposes we focus on education and labor market activities. The state space distinguishes the following states:

0	not working
1	full-time work
2	part-time work (< 35 hours per week)
3	unemployed
4	military service
5	education
6	vocational training
−1	missing information

For each individual, $i = 1, \ldots, 1106$, we then get a sequence

$$(y_{i,0}, y_{i,1}, \ldots, y_{i,142})$$

with y_{it} the state in month t. Of course, there is a certain amount of missing values. In fact, for seven individuals the sequences consist of only missing values. Missing values in the remaining sequences belong to one of the following types:

1. Missing values at the beginning of a sequence do not pose a special problem but only mean that, for some individuals, observation does not begin with the 15th birthday but some months later. In our sample, for 99 % of the respondents the observation periods does not begin

[2] See Blossfeld and Rohwer (1995) for a discussion of this „causal approach“ to modeling interdependent processes.

[3] There is a relatively small amount of attrition that will be shown below.

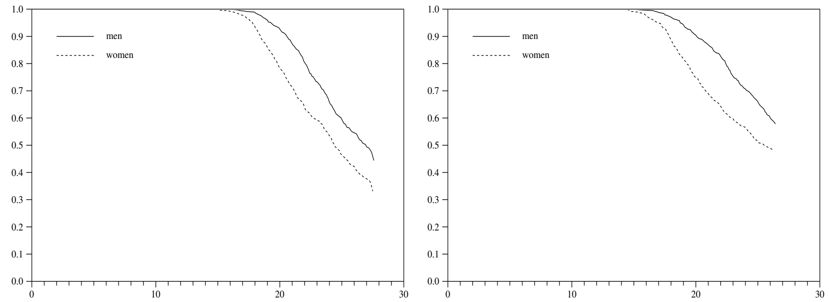

Fig. 2.1 Distribution (survivor function) of age at first marriage (left) and age at first child (right), for men and women born 1964.

later than the 16th birthday.

2. Also missing values at the end of a sequence do not pose a special problem and only mean that, for some individuals, the observation period ends before age 27. This will be illustrated below, see Figures 3.1 and 3.2.
3. There remains the problem of „internal gaps“, that is, missing values in between two valid states. This obviously creates specific difficulties and, in general, one would need suitable assumptions to fill these gaps. Fortunately, in our sample only 42 sequences have internal gaps and so we have decided to use only sequences without internal gaps (and at least one valid state).

The remaining sample used in the illustrations below consists of 1057 individuals (sequences), 550 men and 507 women.

In all illustrations we use only the simple state space described above, focusing on education and work careers. A more complete investigation should also take family events into account. As shown in Figure 2.1, at the end of our observation period (about age 27), almost two third of the respondents are married and almost one half has a first child.

3 Cross-Sectional Distributions

In order to arrive at informative descriptions of a set of sequences one can follow different routes. A basic distinction is:

• We can investigate how the *state distribution*, that is, the distribution of individual units over the state space, evolves in time. This provides a compact description of the aggregate distributions but is basically a

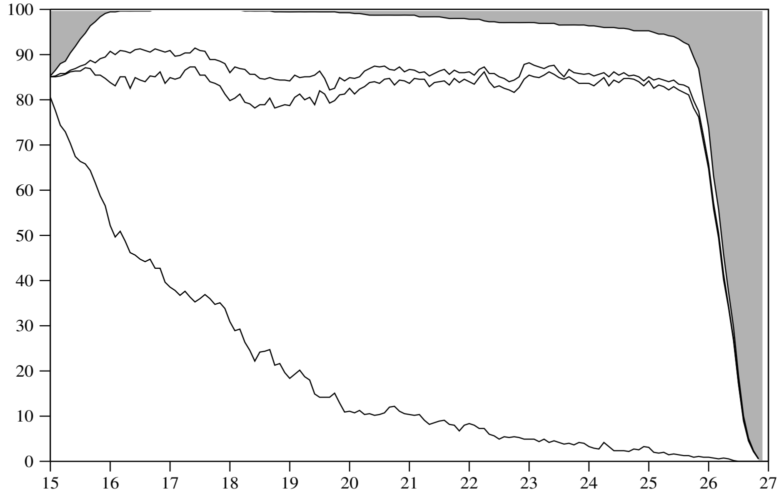

Fig. 3.1 Evolution of state distribution for 550 men. From bottom to top: in education (incl. vocational training), working, unemployed, not working, and missing information.

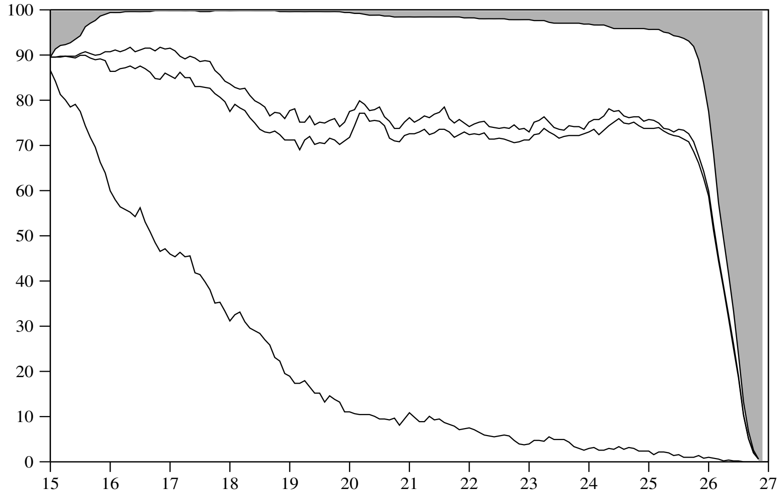

Fig. 3.2 Evolution of state distribution for 507 women. From bottom to top: in education (incl. vocational training), working, unemployed, not working, and missing information.

cross-sectional approach.

- Alternatively, we can be interested in getting information about the behavior of the individual sequences. This will be called an *individual-level approach.*

In this section we illustrate the first approach. In order to provide a general impression, we use an aggregated state space as follows:

0	not working
1	full-time or part-time work, or military service
3	unemployed
5	education or vocational training
−1	missing information

Figures 3.1 and 3.2 show the resulting state distributions for men and women during the observation period (age 15 to 27). The grey-scaled regions indicate the missing observations. We see that there are very little missing values at the beginning, and most sample members can be observed until age 26.

One also sees that the picture is surprisingly similar for men and women. Women are somewhat more often in the non-working state. But, as will be shown in the next section, this is only a gradual difference.

As shown by the two figures, plotting a sequence of cross-sectional state distributions provides a compact view of the structure of a sample of sequences. However, the information is actually very limited and, in fact, might be misleading. As the figures suggest: at the beginning most people are in education or vocational training and then gradually change to some other state, mainly working. However, all our states are repeatable and the figures do not tell us anything about re-entering a previously exited state. The figures do not tell anything about individual mobility between different states and, in fact, there is a risk of interpreting the different regions of the plots as indicating an evolution of „groups of people“.[4] To avoid this risk one needs an individual-level approach.

[4] It seems really difficult to avoid interpreting a sequence of state distribution as showing the development of life courses. For instance, Blossfeld et al. (1993, p. 117) interpret a plot of state distributions similar to Figure 3.1 as showing how „a cohort gradually left the general educational system.“ The problem is that the same sequence of cross-sectional distributions is compatible with a broad variety of different individual careers implying that we cannot draw reliable conclusions about individual careers. And consequently, we cannot draw reliable conclusions about cohorts, given that we should be able to translate statements about cohorts into statements about their individual members.

4 Longitudinal Classifications

The problem indicated in the last section can be generalized: How to find sensible classifications when membership in the classes can change? Given that classifying people into different categories is a main business of sociologists, this is obviously an important problem.

Gershuny (1993), for instance, was interested in the development of occupational careers in Britain and, in order to characterize these careers, he classified his sample members according to the occupation they had at a specific point in time (the year 1986, in his application). He then described the resulting occupational classes by using information about the occupational careers of their members. In particular, he tried to describe these classes by „the proportions of the aggregate working life-times of those in each 1986 occupation who have been in those occupations throughout their careers ('immobile'), and the proportions spent in those occupations by people who have had some other work at some point“ (Gershuny 1993, p. 156). It is questionable, however, whether this construction of occupational classes makes sense. Given that there is some substantive amount of occupational mobility, classifying people according to their occupation at some specific point in time seems highly arbitrary.

In our view, it is generally no good idea to classify people according to their characteristics at only one single point in time. If one wants to define (longitudinal) classes based on time-varying characteristics, one must take into account that people can enter and leave the class, often repeatedly.[5] In a first step, the class should be defined as consisting of all people who were a member of the class at some point in their lives. It is simple, then, to construct a class stability indicator. For each individual i, let D_i denote the potential time that this individual could have been a member of the class, say C, and let D_i^c denote the time during which individual i actually has been a member of C. Then D_i^c/D_i indicates the degree of class membership for this individual; and the distribution of this indicator shows the degree of class stability over time.[6]

[5] Since sociological research is based more and more on longitudinal data, many researchers have recognized this problem; see, e.g., Myles et al. 1993, p. 175. There remains, however, a lack of suitable methods to assess the problem empirically.

[6] One should be aware of the shape of this distribution. In most applications, we will not be able to make a clear distinction between „stable classes“ and „transitory pseudo-classes“. For example (Myles et al. 1993, p. 175): „Traditionally, most male blue-collar workers got their jobs as young men and spent all or a good part of their lives in the same job or circulating between a limited set of similar jobs. [...] The

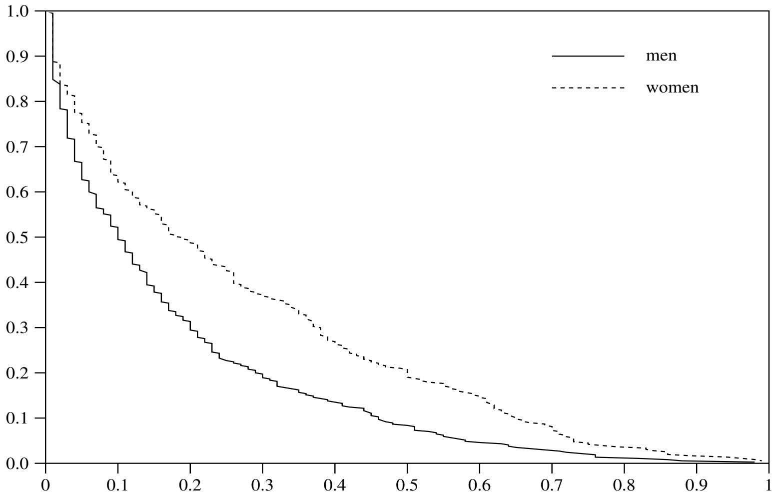

Fig. 4.1 Distribution (survivor functions) of group membership indicator for the „not working“ group. Indicator runs from 0 to 1 on the abscissa.

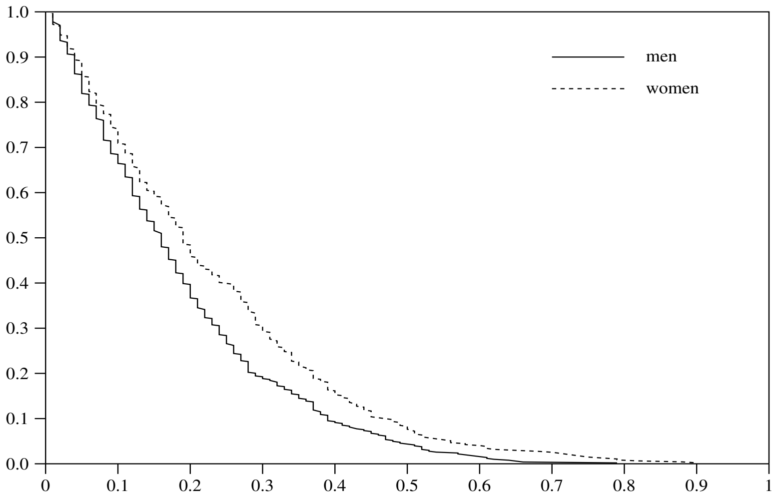

Fig. 4.2 Distribution (survivor functions) of group membership indicator for the „working part-time“ group. Indicator runs from 0 to 1 on the abscissa.

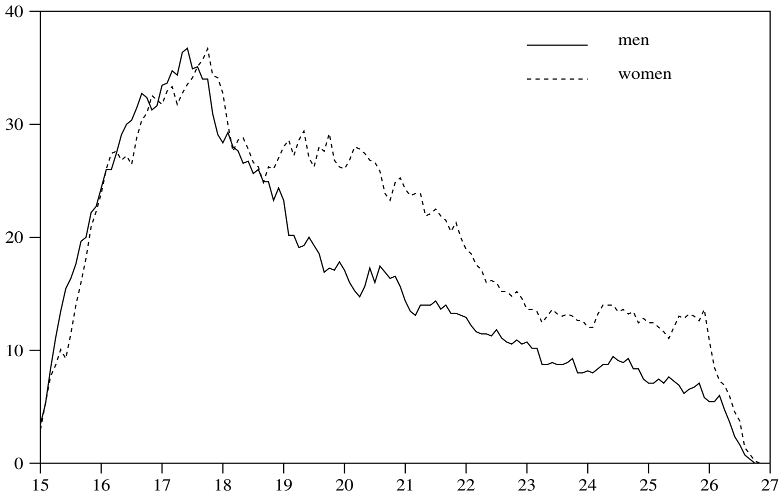

Fig. 4.3 Probability (frequency) of belonging to the „group of part-time working people“ as a function of age (abscissa).

To illustrate this idea, we use the group of people who are „not working“ as suggested by Figures 3.1 and 3.2. For each individual, we calculate the proportion of time (observation period) where the individual belongs to this group. Figure 4.1 shows the distribution, separately for men and women. While women obviously belong somewhat longer to this group compared with men, the plot clearly shows that, neither for men nor for women, there is a stable group membership. Without further qualification, the „group of non-working people“ seems to be an ill-defined concept.

Figure 4.2 shows the distribution of a group membership indicator for the „group of part-time working people“. The conclusion is basically the same and, in this example, there seems to be almost no difference between men and women.

However, this seemingly obvious conclusion provides a starting point for discussing another limitation of longitudinal classifications. While Figure 4.2 seems to suggest that men's and women's careers are quite similar with respect to part-time work, this is actually not the case.

question then is whether the post-industrial working class is like this. Or are the low-skilled, low-wage service jobs 'stop-gap' jobs, places which people 'pass through' on their way to somewhere else?“ Most probably, this is not an alternative.

This impression is simply the result of ignoring the distribution of part-time work in different ages. We see this when calculating the probability (frequency) of belonging to the „group of part-time working people“ as a function of time (age). This is shown in Figure 4.3. The plot clearly shows that the probability of part-time work is similar for men and women only until about age 19, that is, in a period with mainly „stop-gap“ jobs; but afterwards women have a significantly higher probability of working part-time.

As this examples demonstrates, whenever defining longitudinal classifications, this should be supplemented by (a) an investigation of the stability of group membership over time, and (b) an investigation of the probability of belonging to a group as a function of time (age).

5 Characterizing Individual Sequences

We now turn to individual-level approaches to describe life courses. A simple approach calculates the proportion of time (observation period) each individual spent in the possible states. Figure 5.1 shows distribution functions for these proportions. If a functions begins with a value less than 1, the remaining proportion of sample members never experienced the corresponding state.

Another approach is based on the view of life courses as sequences of events. It seems possible, then, to characterize each individual sequence by reporting dates and frequencies for the occurrence of specific events. The distributions of these indicators can then be used to compare, for instance, life courses of men and women, or across different birth cohorts.[7] However, while this approach is straightforward for simple, non-repeatable events like first marriage, there are other events which are more difficult to describe: (a) repeatable events, like becoming unemployed, and (b) complex events as, for instance, the transition from education and vocational training to work.

Repeatable events. A useful approach to describing the occurrence of repeatable events is to calculate the probabilities (or rates) for the occurrence of such events as a function of time. Based on a sequence representation of life courses this can be done easily. One just needs to calculate, for each point in time, a risk set containing the individuals who might experience the event, and the number of individuals who actually experienced the event. Dividing the latter by the former provides a simple

[7] See Mayer (1993) for an application of this approach to cohort comparison.

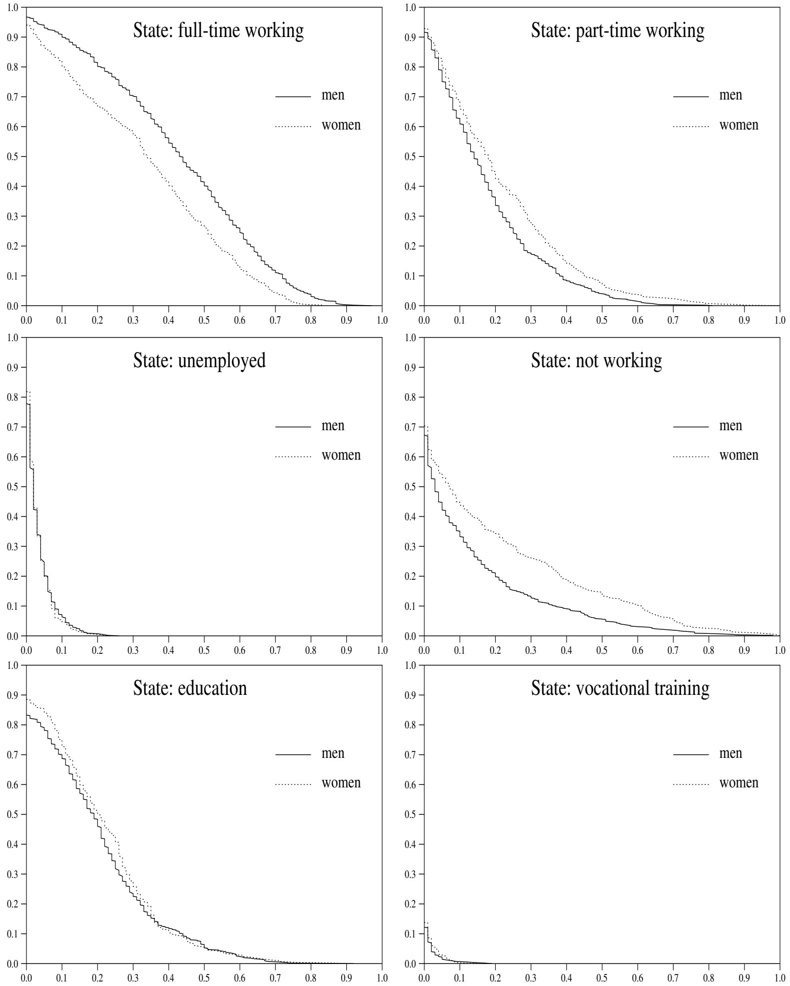

Fig. 5.1 Distribution (survivor functions) for proportion of time (observation period) spent in different states.

estimate of the probability.

This calculation can be done separately for each type of event that seems interesting. In a more general approach, one can also calculate the probability for the occurrence of any event, that is, for any change in the current state. This would then provide an overall measure for the degree

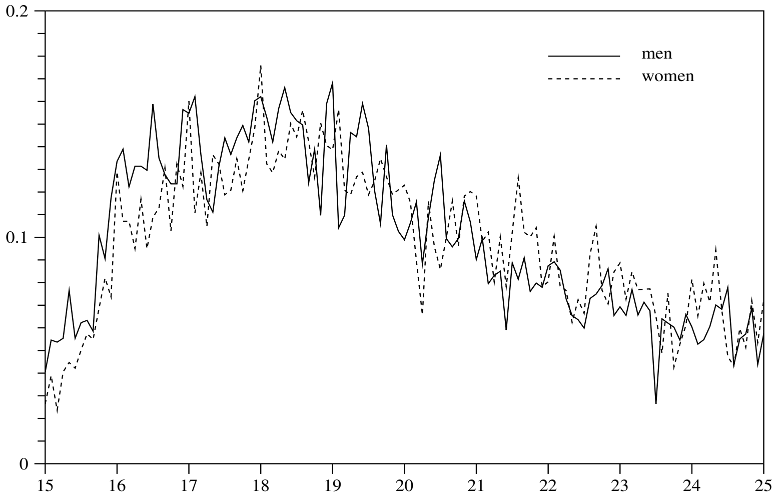

Fig. 5.2 Probability (frequency) of the occurrence of any events, based on a state space with seven different states.

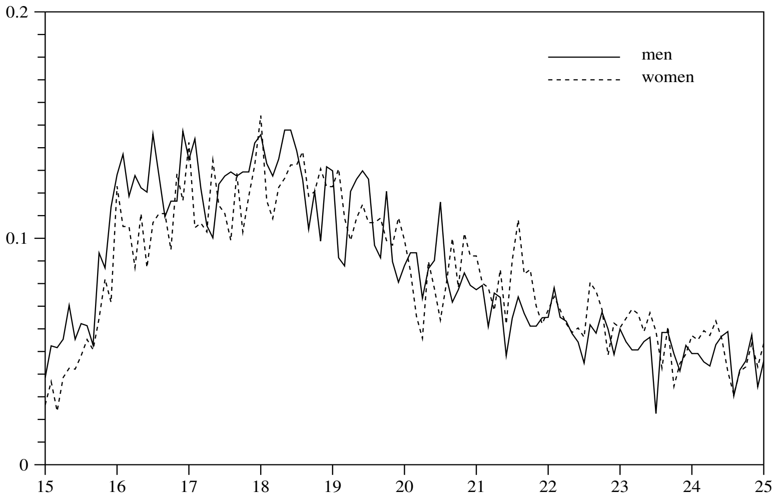

Fig. 5.3 Probability (frequency) of the occurrence of any events, based on an aggregated state space with four different states.

of mobility, as a function of time. We take this latter approach for an illustration. Figure 5.2 shows how the probability (frequency) of events depends on age, based on our standard state space with seven different states. As just mentioned, this can be viewed as an overall indicator of time-varying mobility. In this example, mobility reaches its highest values in the period where young men and women are moving from education to vocational training and work. Interestingly, the degree of mobility is basically the same for men and women.

In general, the degree of mobility will depend on the number of different states and, consequently, the number of possible events. However, in this example we find very little difference when using the aggregated state space that only distinguishes four different states (see the definition in Section 3). Comparing Figure 5.3 with Figure 5.2 shows almost the same amount of mobility.

Complex events. We use the word „complex events“ to denote events which do not necessarily happen in a single time unit but may take the form of a transition period. A typical example is the transition from education to work where we assume that a person is „mainly in education“, for a certain period, and then „mainly working“ for another period. However, there might be another period in between these two situations where the individual is in a certain mix of different states.

We will use this example for an illustration. We first define, somewhat arbitrarily, two time points, separately for each individual in our sample:

- $t_{i,1}$ the earliest time point t such that individual i is in education at t and is not in education for at least 12 months following t
- $t_{i,2}$ the earliest time point $t \geq t_{i,1}$ such that individual i is not working at t and is working for at least 12 months following t

We can then use $t_{i,2} - t_{i,1}$ as indicating the duration of the transition period from education to work. Table 5.1 shows the distribution. For 125 men and 102 women we cannot find a transition from education to work. 212 men and 172 women change immediately from education to a period of mainly working. For 213 men and 233 women we find a transition period having a duration of at least one month but may extend to several years.

Figures 5.4 and 5.5 show the distribution of the length of the transition period for the men and women, respectively, where this length is at least one month. In addition, the figures show the kinds of activities (states)

Tab. 5.1 Transition period from education to work

Transition period	Men	Women
no event	125	102
immediate transition	212	172
1 – 12 months	134	131
13 – 24 months	23	36
more than 24 months	56	66

during this transition period; from bottom to top: not working, education (incl. vocational training), military service, unemployed, and working.

Searching for Patterns. The idea underlying our description of a transition period from education to work can be generalized into a general search for patterns. Given a sequence of states, a pattern is simply a predefined subsequence of states. For instance, in our sequence data, the pattern

$$p = (5, 5, 5, 5, 5, 5, 6, 6, 6, 1, 1, 1, 1, 1, 1)$$

would consist of six months in education, followed first by three months in vocational training and then by six month in full-time work. Of course, it is quite unlikely to find exactly this pattern in our sample; in fact, it does not occur in any of our sequences. The idea of pattern matching can be saved, however. One possibility is to use only partly specified patterns, for instance

$$p = (5, *, 6, *, 1)$$

This would mean: at least one month in education, followed by an arbitrary (possibly empty) sequence of states, followed by at least one month in vocational training, followed again by an arbitrary (possibly empty) sequence of states, and finally at least one month working. Then we find already 95 sequences containing this pattern at least once. Table 5.2 presents a few examples.

Another approach would be to search for *approximate* pattern matches. This would require the definition of a similarity measure for comparing patterns with sequences and to find those parts of a sequence which are most similar to the predefined pattern. However, we have not yet explored whether this could lead to a useful method for describing sequences.

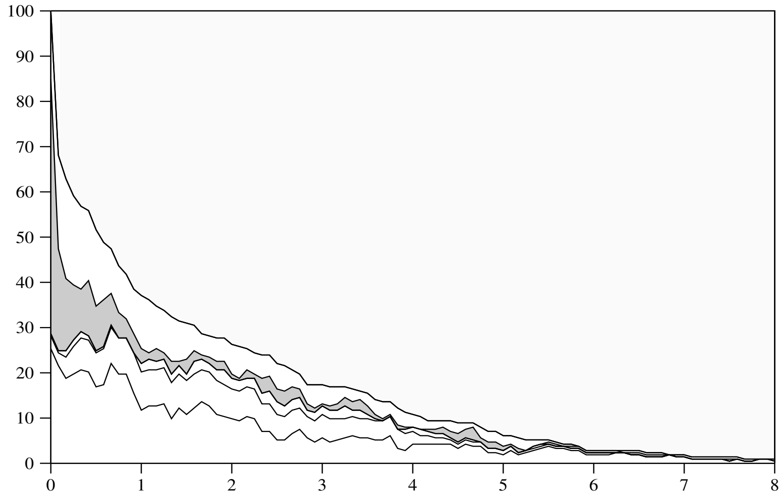

Fig. 5.4 Distribution (survivor function) of the length of the transition period from education to work for 213 men; and distribution of states during this transition periods.

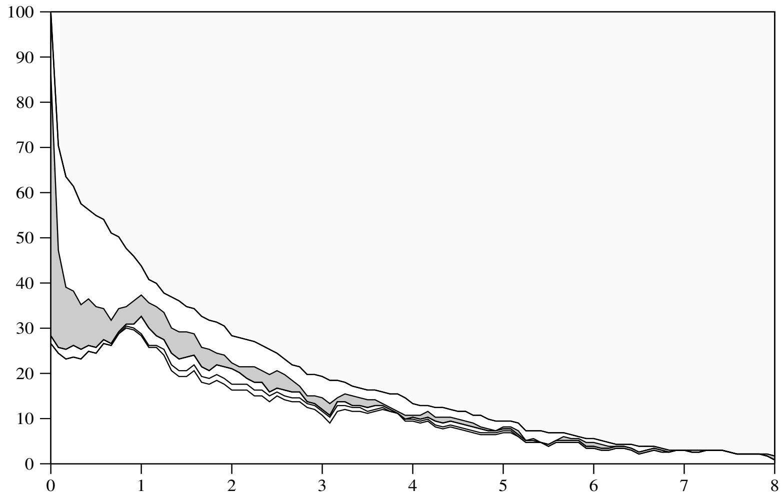

Fig. 5.5 Distribution (survivor function) of the length of the transition period from education to work for 233 women; and distribution of states during this transition periods.

Tab. 5.2 Occurrence of patterns in sequences

Pattern	Men yes	Men no	Women yes	Women no
EDU,*,VT,*,WORK	44	506	51	456
EDU/VT,*,WORK,*,EDU/VT	364	186	324	183
WORK,*,VT,*,WORK	8	542	11	496
WORK,*,EDU/VT,*,WORK	375	175	332	175
WORK,*,UNEMP,*,WORK	356	194	354	153

6 Searching for Typical Careers

In the literature surrounding the POLIS project, the notion of „life course regimes“ plays a central role. While it would be difficult to provide a clear definition of this concept, it certainly implies the idea of typical careers and the assumption that real people's life courses are in some sense similar to the typical careers constructed by sociologists. Are there any chances to investigate this idea empirically?

The problem is obviously complex. Given a state space with q different states, there are q^l different sequences of length l. This number gets easily very large, exceeding not only normal sample sizes but also the size of all known populations. In our sample, $q = 7$ and $l = 143$ resulting in about 10^{123} different sequences. Trying to find the most often occurring sequences is certainly not a good idea. In fact, in our sample, not two sequences are identical.

Whatever the appropriate way to defining typical careers, one cannot expect to find exactly this career very often in an empirically given sample of sequences. There are two kinds of strategies to cope with this problem. First, one can follow a strategy of aggregation. One possibility would be to consider broader time units, say years, instead of months and the predominating states during these broader time units. This would then substantially diminish the degree of heterogeneity in the sequences and may finally lead to more easily comprehensible frequency distributions. Another possibility would be to only consider the ordering of states, while ignoring their duration. This has been tried, for instance, by Rindfuss et al. (1987) and Berger et al. (1993). In our view, these aggregating approaches are questionable because they count states without considering their duration, or simply ignore potentially important events. Moreover, what is left out by the aggregation procedure is practically not controllable.

An alternative strategy tries to systematically face the problem of heterogeneity in each given sample of careers. The basic requirement is then to define, in a first step, a measure of proximity that can be used to assess the degree of similarity or dissimilarity between careers. There are basically two different approaches.

a) The first approach can be called cross-sectional. Given two sequences, $y_i = (y_{it})$ and $y_j = (y_{jt})$ (for two individuals, i and j), the approach begins by defining a cross-sectional distance measure $d(y_{it}, y_{jt})$ that allows to calculate the distance of the two sequences for each point in time, t. The overall distance of the two sequences is then calculated by aggregating the cross-sectional distances over a certain range of the time axis. This approach has been used, for instance, by Buchmann and Sacchi (1995) to find a classification of occupational careers. Also, when using correspondence analysis to find „structure“ in a set of sequences, the underlying distance measure is basically cross-sectional; for an application of this approach see, e.g., Martens (1994).

This cross-sectional approach has two drawbacks. First, it can only be used to compare sequences of equal length. Second and more important, it cannot appropriately cope with situations where only the timing of events is different in the sequences. Assume, for instance, two sequences which both have one month of unemployment and which are identical in all respects except that the unemployment occurs one month later in the second sequence. A sensible distance measure should make both sequences quite similar, but the cross-sectional approach would count a difference in two months.

b) To avoid the shortcomings of the cross-sectional approach, some sociologists have tried an optimal matching approach to comparing sequences.[8] This approach provides the possibility to compare individual sequences simultaneously at different points in time. Although there isn't much experience with this approach in sociological research, we should expect it to provide proximity measures which are better suited to the dynamical nature of life courses.

While the availability of a suitable proximity measure for sequences is a basic prerequisite, it does not automatically provide a definition of typical careers. Again, different approaches are possible.

- One can simply define idealized careers, based on whatever evidence

[8] See, e.g., Abbott 1983, 1995; Abbott and Hrycak 1990; Wing 1995; Stovel et al. 1996.

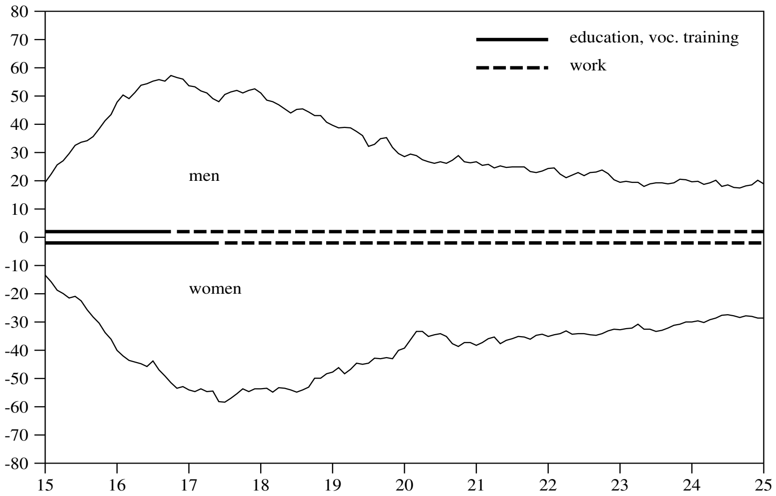

Fig. 6.1 Typical careers constructed by using the most frequent state in each month. In addition: percentage of sequences that were not in the typical state in the current month.

and prejudices are available. Then one can investigate the degree of similarity between these idealized careers and the sequences in a given sample. Based on the distribution of the resulting similarity indices, one might be able to answer the question whether the predefined careers provide a useful representation of the empirically given sequences.

• Another approach begins with calculating a similarity index for each pair of sequences in the given sample. The resulting distance matrix can then be used as input for some clustering procedure in order to find a set of sequence clusters. Finally, one can try to represent each cluster by a typical sequence.

In the remainder of this section we shortly illustrate the first approach. The problem of classifying careers will be discussed in the next section.

One way of defining a (potentially) typical career is based on using the most frequent state in each time unit (month). If we follow this way with our sample of sequences, we find the sequences shown in Figure 6.1. For men, the sequence is 22 months mainly in education, then 120 months mainly working; for women it is 30 months mainly in education,

then 113 months mainly working.[9] While the plot looks fine, the important question is, of course, whether and to what degree the individual careers actually follow these sequences. Are they, in some sense, „typical careers“?

To approach this question, Figure 6.1 also shows the proportion of sequences that are not in the most frequent state in each current month. For instance, at age 17, more than 50 % of the sequences are not in their most frequent state; and at least 20 % of the sequences are always, i.e. in each month, different from the most frequent state. It is questionable, therefore, whether the idealized careers shown in Figure 6.1 provide a useful summary of the actual variety of careers in our sample. In particular, while these idealized careers suggest a clear transition from education to work, we have already seen in section 5 that there is, in fact, a highly complex transition period characterized by an extremely high amount of individual heterogeneity.

Moreover, the method used in Figure 6.1 to assess the amount of non-typical careers is basically cross-sectional and somewhat misleading. We have simply calculated the percentage of sequences that were not in the most frequent state for each month separately. To investigate the question correctly, we need to calculate an explicitly defined distance between each individual sequence and the idealized careers. To illustrate this approach, we use optimal matching distances in their canonical definition, that is, insertion and deletion weights are both 1, and substitution weights are 2.[10] Each comparison between a sequence and an idealized career uses the longest common sequence length, and the resulting distance is normalized by dividing through the common sequence length. The resulting distances can then vary in the range 0 (identical) to 2 (most dissimilar).

Figure 6.2 shows the distribution of these distance measures and suggests the conclusion that most sequences in our sample are not very similar to the idealized careers shown in Figure 6.1. It is, of course, difficult to justify any threshold such that a distance below the threshold can reasonably being interpreted as „being similar to the idealized career“. A potentially useful method to provide additional evidence is to calculate distances simultaneously with respect to two or more reference sequences. Figure 6.3 illustrates this idea by showing simultaneously the distance of men's sequences to the idealized sequence (Y-axis) and a zero

[9] The construction is based on the aggregated state space described in Section 3.

[10] This amounts to using the length of the longest common substring as a measure of similarity between sequences; see, e.g., Kruskal, 1983.

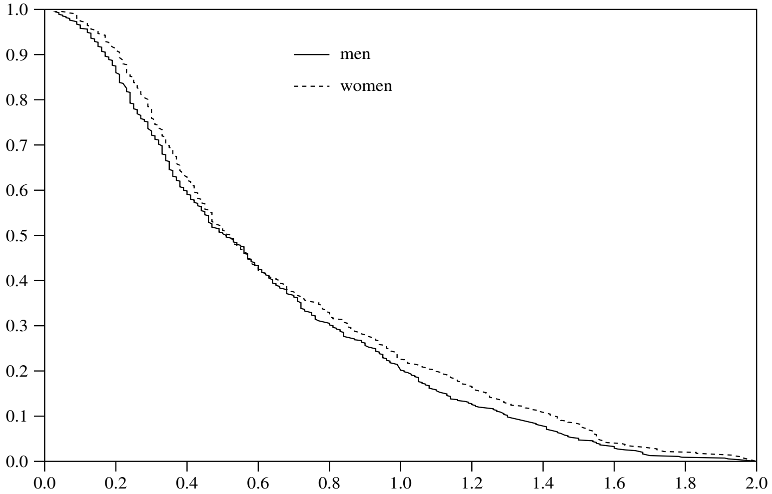

Fig. 6.2 Distribution (survivor function) of the normalized distances between individual sequences and the typical sequences shown in Figure 6.1.

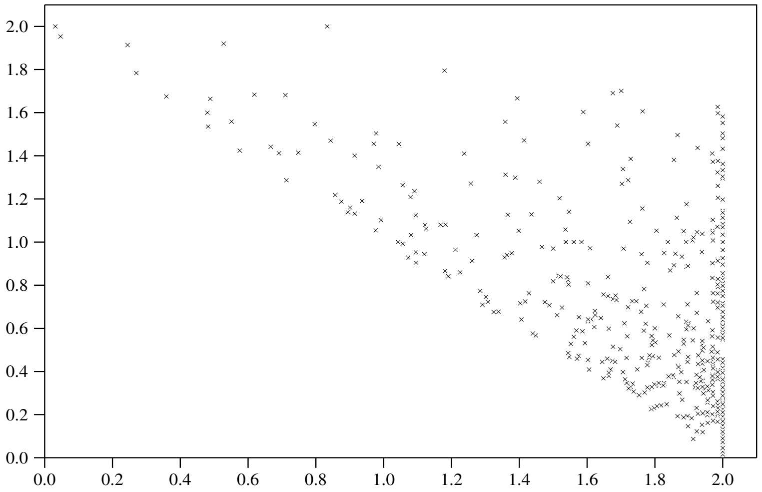

Fig. 6.2 Scattergram of men's sequences. X-axis: distance to a zero sequence (always not working), Y-axis: distance to the typical sequence shown in Figure 6.1.

sequence consisting of always not working (X-axis). We see a substantial part of the sequences being similar to the idealized career and not similar to the zero sequence. However, the main conclusion is again that one single sequence cannot represent the broad diversity of actual sequences.

7 Classifying Careers

Given that we cannot expect to find a single typical career which appropriately summarizes a given sample of sequences, it seems reasonable to search for a small set of different career types. Given also that we do not have good a priori reasons for defining these career types, we could try some clustering procedure to empirically investigate the question.

To illustrate this approach we use again the optimal matching distances introduced in the previous section. Instead of comparing each sequence with a predefined idealized career, we now calculate a distance matrix that provides a distance between each pair of sequences. For the illustration we begin with using the 550 sequences of men, resulting in a symmetrical (550, 550) matrix with (up to) 150,975 different distances in its lower triangle.

Several different algorithms are available for clustering a distance matrix. We do not try to find an optimal algorithm but choose a particularly simple one to discuss some general problems. The algorithm is a binary split procedure which, beginning with the whole sample, sequentially tries to find optimal subclasses. The first steps of the split procedure are shown in Table 7.1, the resulting classification tree is shown in Figure 7.1.[11]

The algorithm begins with the whole sample consisting of 550 sequences for men; this is level 0 and the class ID is 1. It then finds two sequences which are maximally different to become the seed points for two subclasses on level 1. All sequences which are more similar to the first than to the second of these sequences will become members of the first subclass (class ID 2), the remaining sequences become members of the second subclass (class ID 3). In our example, the first subclass has 452, the second subclass has 98 members.

The procedure is then repeated for each previously created subclass having at least two members. The algorithm finds two maximally different elements (sequences) to become the seed points for two further

[11] For plotting this tree we used the radial drawing algorithm proposed by Barthelemy and Guenoche 1991, p. 27.

Tab. 7.1 First steps of binary split procedure

Level	Classes	Class ID	Elements	Max Distance	Mean Distance
0	1	1	550	2.00	0.83
1	2	2	452	2.00	0.73
1	2	3	98	2.00	0.85
2	4	4	350	1.77	0.52
2	4	5	102	1.96	0.94
2	4	6	72	1.83	0.72
2	4	7	26	1.63	0.85
3	8	8	170	1.28	0.38
3	8	9	180	1.43	0.48
3	8	10	38	1.50	0.80
3	8	11	64	1.43	0.76
3	8	12	40	1.55	0.67
3	8	13	32	1.08	0.48
3	8	14	11	1.34	0.78
3	8	15	15	1.05	0.63
4	16	16	163	1.08	0.33
4	16	17	7	0.92	0.51
4	16	18	173	1.12	0.47
4	16	19	7	0.83	0.63
4	16	20	27	1.24	0.63
4	16	21	11	1.28	0.86
4	16	22	35	1.22	0.72
4	16	23	29	1.10	0.63
4	16	24	28	0.92	0.54
4	16	25	12	1.17	0.58
4	16	26	3	0.81	0.64
4	16	27	29	1.00	0.40
4	16	28	5	0.94	0.71
4	16	29	6	1.00	0.63
4	16	30	8	0.85	0.65
4	16	31	7	0.91	0.46

subclasses. Table 7.1 documents this process up to level 4 with 16 subclasses.

The question remains, of course, where to stop the procedure. Figure 7.1 shows the final classification tree but provides no clear idea about the number of classes. (A similar classification tree for women is shown in Figure 7.2.) The plot seems to suggest that we should distinguish at least 11 different classes. But we need additional information to assess

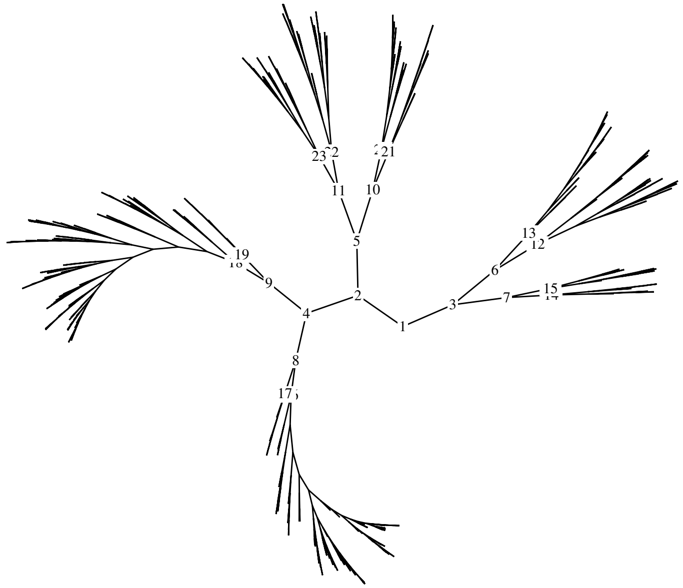

Fig. 7.1 Classification tree resulting from a binary split procedure applied to optimal matching distances between sequences of 550 men.

their homogeneity. Table 7.1 records the maximal distance between two sequences in each of the subclasses but this is not a very good indicator of subclass homogeneity. A somewhat better measure is provided by the mean of all pairwise distances in a subclass. Using this measure, also shown in Table 7.1, seems to suggest that even 16 classes may not be sufficient to approach an acceptable degree of homogeneity.

Assessing Longitudinal Diversity. The problem of finding useful classifications of sequences (careers) can certainly not be solved on purely formal and statistical grounds. We should recognize, however, that clustering sequences brings us back to the general problem, already discussed

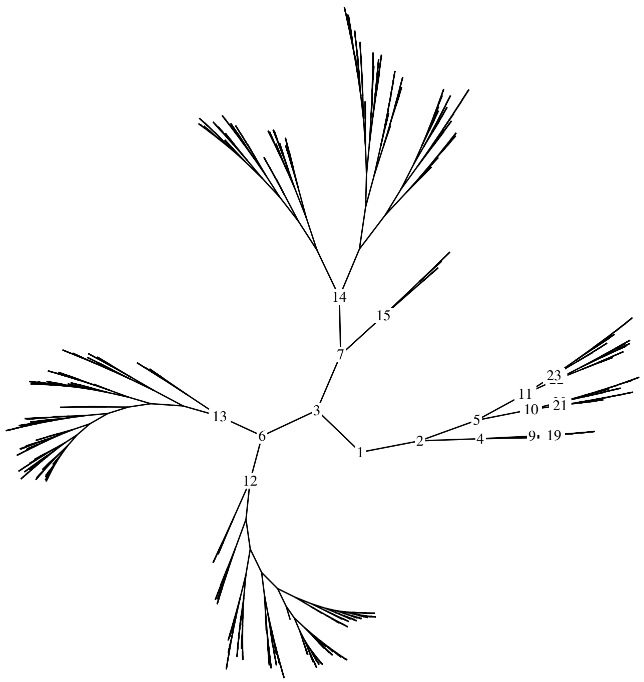

Fig. 7.2 Classification tree resulting from a binary split procedure applied to optimal matching distances between sequences of 507 women.

in section 4, how to define *longitudinal* classifications. The approach taken in section 4 was based on a fundamental assumption: that any reasonable classification should allow people to change their class membership. This assumption provides the opportunity to define classes on theoretical grounds and then investigate class membership as a function of time. If classifying whole sequences, this assumption is no longer valid and homogeneity of classes becomes an essentially problematic claim.

An important part of the problem follows from the fact that classifying whole careers contradicts, in a sense, a dynamical view of life courses as sequentially developing in time. We have no good idea how to cope with this contradiction. In any case, it seems worthwhile also directly

investigating the question how the diversity of life courses evolves over time (age).

There are basically two different approaches. First, a cross-sectional approach where we calculate an indicator of diversity (inequality) for each time unit separately and then plot this indicator as a function of time. This approach is directly related to the sequences of state distributions discussed in section 3. To illustrate this approach, we calculate separately for each month the entropy of these distributions which can be interpreted as a measure of homogeneity.[12] Using our basic state space with seven different states, the entropy can vary in the range from 0 (all sequences are in the same state) up to $\log(7) = 1.95$ (all sequences are equally distributed across the different states). Figure 7.3 shows the entropy as a function of time. It begins with a very low values since most sample members are then in the same state (education). The sequences then become rapidly more differentiated and the entropy reaches almost its maximum at about 19 years; and then the entropy slowly decreases.

Cross-sectional measures of diversity are questionable, however, since they assume that life courses do not have a memory. These measures only take into account the currently realized state and completely ignore what has happened before. Taking this objection seriously, we should sequentially compare our sequences. This can be done by generalizing the optimal matching approach discussed above. Instead only calculating a single distance for each two sequences, we now do this sequentially for each month t, $d_{ij,t}$ being the distance of the sequences of individuals i and j up to the tth month. The result is a sequence of distance matrices, a separate distance matrix for each month reflecting the inequality structure of the sequences up to that month. Since we do not yet have a good idea how to explore the structure of this sequence of distance matrices, Figure 7.4 only shows the mean values of the distances as a function of time.

8 Concluding Remarks

As mentioned at the beginning, statistical methods for describing life courses are not well developed yet. The preceding sections reflect this state of affairs. While it seems possible to find useful descriptions for certain aspects of life courses, we soon reach limits when trying to describe whole trajectories. This is partly a consequence of the inherent

[12] The entropy in month t is calculated as $E_t = \Sigma_j p_{jt} \log(p_{jt})$ where p_{jt} is the proportion of sequences being in state j in month t.

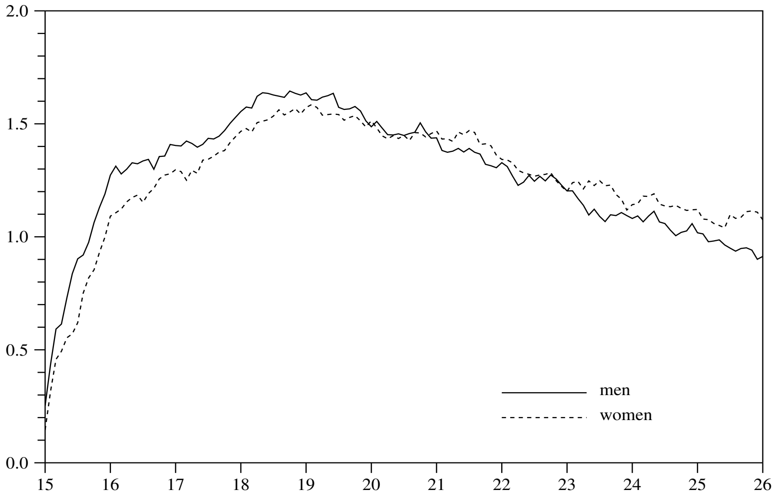

Fig. 7.3 Entropy (homogeneity) of state distributions as a function of time, based on seven different states.

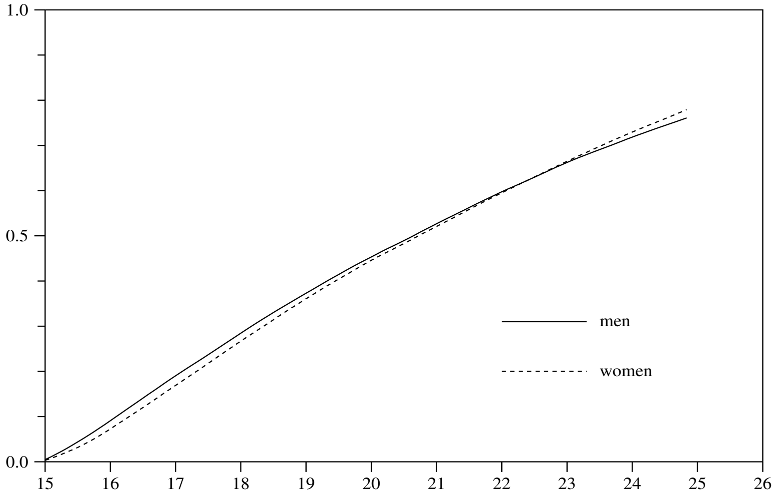

Fig. 7.4 Mean values of optimal matching distances $d_{ij,t}$, calculated for $t = 1, \ldots, 120$.

complexity of life courses. Possibly more important is the requirement that appropriate methods for describing life courses should always follow the view that life courses evolve sequentially in time. Our preliminary conclusions are as follows.

• When trying to describe life courses we should avoid any form of cross-sectional approach and, in particular, we should avoid aggregating individuals on a cross-sectional basis. Appropriate descriptions should always be interpretable in terms of individual trajectories.

• This does not exclude the possibility of useful classifications. However, we should make an important distinction between classifying individuals and classifying possible states of individual life courses. Taking the life course approach seriously, only the latter type of classification is potentially sensible. Then, whenever trying to establish some classification, this should be supplemented by an investigation of class membership as a function of time.

• Trying to classify whole life courses seems to be an essentially problematic endeavor. While we easily admit that more sophisticated methods can be developed to find potentially useful classifications of whole trajectories, there remains a basic contradiction with a dynamical view of sequentially evolving life courses. In our view, the basic task is to reach an appropriate description of the development of diversity (and inequality) of life courses over time. Trying to find better methods for classifying whole trajectories should be seen as a secondary task.

References

Abbott, A. (1983). Sequences of Social Events: Concepts and Methods for the Analysis of Order in Social Processes. Historical Methods 16, 129 – 147.

Abbott, A., Hrycak, A. (1990). Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musician's Careers. American Journal of Sociology 96, 144 – 185.

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

Barthelemy, J.-P., Guenoche, A. (1991). Trees and Proximity Representations. New York: Wiley.

Berger, P. A., Steinmüller, P., Sopp, P. (1993). Differentiation of Life-Courses? Changing Patterns of Labour-Market Sequences in West Germany. European Sociological Review 9, 43 – 65.

Blossfeld, H.-P., Rohwer, G. (1995). Techniques of Event History Modeling. Mahwah, NJ: Lawrence Erlbaum.

Blossfeld, H.-P., Giannelli, G., Mayer, K. U. (1933). Is There a New Service Proletariat? The Tertiary Sector and Social Inequality in Germany. In: G. Esping-Andersen (ed.): Changing Classes: Stratification and Mobility in Post-industrial Societies, pp. 109 – 135. Newbury Park: Sage.

Buchmann, M., Sacchi, S. (1995). Mehrdimensionale Klassifikation beruflicher Verlaufsdaten. Eine Anwendung auf Berufslaufbahnen zweier Schweizer Geburtskohorten. Kölner Zeitschrift fr Soziologie und Sozialpsychologie 47, 413 – 442.

Center for Human Resource Research (1994). NLS Handbook 1994. The National Longitudinal Surveys. Columbus, Ohio: The Ohio State University.

Gershuny, J. (1993). Post-industrial Career Structures in Britain. In: G. Esping-Andersen (ed.): Changing Classes: Stratification and Mobility in Post-industrial Societies, pp. 136 – 170. Newbury Park: Sage.

Kruskal, J. B. (1983). An Overview of Sequence Comparison. In: D. Sankoff, J. B. Kruskal: Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison, pp. 1 – 44. Reading: Addison-Wesley.

Martens, B. (1994). Analyzing Event History Data by Cluster Analysis and Multiple Correspondence Analysis. In: M. Greenacre, J. Blasius (eds.): Correspondence Analysis in the Social Sciences, pp. 233 – 251. New York: Academic Press.

Mayer, K. U. (1993). Gesellschaftlicher Wandel, Kohortenungleichheit und Lebensverläufe. In: L. Montada (ed.): Bericht über den 38. Kongreß der Deutschen Gesellschaft für Psychologie in Trier 1992, pp. 73 – 92. Gttingen: Hogrefe.

Myles, J., Picot, G., Wannell, T. (1993). Does Post-industrialism Matter? The Canadian Experience. In: G. Esping-Andersen (ed.): Changing Classes: Stratification and Mobility in Post-industrial Societies, pp. 171 – 194. Newbury Park: Sage.

Rindfuss, R. R., Rosenfeld, R. A., Swicegood, C. G. (1987). Disorder in the Life Course: How Common and Does it Matter? American Sociological Review 52, 785 – 801.

Rohwer, G. (1996). TDA User's Manual, Part IX: Sequence Data. Berlin: Max Planck Institut für Bildungsforschung.

Stovel, K., Savage, M., Bearman, P. (1996). Ascription into Achievement: Models of Career Systems at Lloyds Bank, 1890 – 1970. American Journal of Sociology 102, 358 – 399.

Chan, T. W. (1995). Optimal Matching Analysis: A Methodological Note on Studying Career Mobility. Work and Occupations 22, 467 – 490.