Timing, sequencing, and quantum of life course events: a machine learning approach

Francesco C. Billari
Max Planck Institute for Demographic Research,
Doberaner Str. 114, D-18057 Rostock, Germany
E-mail: billari@demogr.mpg.de

Johannes Fürnkranz
Austrian Research Institute for Artificial Intelligence
Schottengasse 3, A-1010 Wien, Austria
E-mail: juffi@ai.univie.ac.at

Alexia Prskawetz
Max Planck Institute for Demographic Research,
Doberaner Str. 114, D-18057 Rostock, Germany
E-mail: fuernkranz@demogr.mpg.de

Abstract

In this methodological paper we discuss and apply machine learning techniques, a core research area in the artificial intelligence literature, to analyse simultaneously timing, sequencing, and quantum of life course events from a comparative perspective. We outline the need for techniques which allow the adoption of a holistic approach to the analysis of life courses, illustrating the specific case of the transition to adulthood. We briefly introduce machine learning algorithms to build decision trees and rule sets and then apply such algorithms to delineate the key features which distinguish Austrian and Italian pathways to adulthood, using Fertility and Family Survey data. The key role of sequencing and synchronisation between events emerges clearly from the methodology used.

Keywords: life course, event history, data mining, machine learning, transition to adulthood.

1 Introduction

Demographers are mostly concerned with the study of major events that shape people's lives such as births, deaths, migrations, and the formation and dissolution of households and families. The life course approach, which has been the theoretical framework behind many recent studies, sees above all the demographic trajectories and transitions of individuals as a series of parallel trajectories which can be embedded in other trajectories [31]. These can be marked by demographic events in the family or the residential spheres, or by events which are thought to have an influence on demographic behavior (educational and labor market careers are among the most important in the latter category). The life course approach to the study of demographic behavior is thus characterised by a holistic point of view. It thus specifically makes sense to conceive of what we might call 'demographic life courses' as—at least partially—the consequences of strategic behavior. That is, individuals have some general ideas about the future development of their lives. Research can then only gain from the use of an approach which takes a holistic point of view on life courses, because decision–making is also partially guided by unitary principles [20]. The problem is that the techniques used so far hardly allow one to take such a holistic perspective.

One of the fields that has attracted increasing interest in the demographic life course literature of the last years is the study of the transition to adulthood. In this field, the main emphasis is on the study of the *timing*, the *sequencing* [17], and sometimes the *quantum* of specific events—which usually happen during early adulthood—for a specific cohort of individuals. These events are normally considered to be indicators of the transition from roles typical of youth to roles typical of adulthood. For the sake of simplicity, the age at which events are experienced is taken as an indicator of the timing, the observed order as an indicator of sequencing, and the observed number of events as an indicator of the quantum. In the latter case (quantum), if one focuses on the transition to adulthood, the main issue is whether an event is experienced at all during the life of an individual. As far as the sequencing is concerned, a critical issue is the simultaneity of events, i.e., the experiencing of events in the same time unit, also known as *synchronisation* [26]. Following a seminal paper by Modell et al. [25], most of the papers studying the transition to adulthood analyze some specific events: leaving formal education, entering the labor market, leaving the parental home, experiencing the first union (sometimes with a differentiation between marriages and consensual unions), and becoming a parent. This approach is not the only one that could be adopted in a study on the transition to adulthood [22], but it provides a widely used framework when one wants to analyze the determinants and study the dynamics of behavior within a society by comparing cohorts, genders, social groups, and/or different societies.

For the study of events, the set of statistical techniques which is now broadly defined as *event history analysis* constitutes undoubtedly one of the principal toolkits of demography (see e.g. [9]). Event history techniques focus on the *time-to-*

event as the dependent variable, and they allow researchers to study very complex interdependencies between events in the life course, also handling unobserved factors underlying these complex interdependencies [21]. Event history analysis does not, however, allow one to adopt a holistic perspective on the life course, that is, to see the set of events that shape the lives of individuals as a coherent set and to compare this set for different individuals or groups of individuals. Event history analysis normally allows for the analysis of the timing and, with some specific assumptions, also of the quantum of events, but it does not allow for the simultaneous study of the sequencing of events. In general, we might say that it is not possible to adopt a holistic perspective using event history analysis, because the life course in its conceptual unit cannot be taken as the variable to be analyzed.

Hence, different techniques have to be envisaged if one wants to look at life courses from a holistic point of view. Such techniques take as their starting point the fact that a representation of life courses based on a sequence of states is equivalent to a representation based on events, as long as events are recorded on a discrete timeline. The *sequence analysis* approach is based on a sequential representation of life courses . It was first applied in the social sciences by Abbott (see the reviews in [1, 2]). In the sequence-based approach, individual life courses are represented using a time-oriented string which contains the states occupied at each point in time (e.g., each month or each year) instead of the events that cause state transitions. The main problem is that the analysis of such data is very complex. One cannot use in the sequence directly as a dependent variable that has to be explained by a statistical model. Indeed, it is very likely that in a sample of individuals, every individual will be characterised by a different sequence. Classification has thus become the major approach to sequence data analysis in the social sciences. Some techniques used in the natural sciences, such as optimal matching, have been used to cluster life courses in different groups. However, it is sometimes difficult to identify the reasons why some individuals are assigned to a specific group, and it is unclear how much this assignment depends on the distance between states, which has to be assigned subjectively by the researcher. Other techniques for the analysis of sequence data are based on monothetic divisive algorithms that result in a classification tree [4]; the latter have been proposed for grouping individuals according to their states at different points in time. While these trees are quite similar to the decision-tree learning techniques used in this paper, so far their use has been restricted to the study of non-repeatable events, and they do not explicitly take into account the order of events.

In this paper, which was written primarily for methodological reasons, we propose a solution to the problem of analyzing simultaneously the timing, the sequencing (including the synchronisation on a monthly time basis), and the quantum of events in life courses. To this aim, we employ some techniques that have been developed in the field of Artificial Intelligence. In particular, we use state-of-the-art machine learning algorithms to detect the basic features (of timing, sequencing, and quantum) that differentiate two groups of life courses. We apply two machine learning algorithms to analyze the transition to adulthood in two European coun-

tries, Austria and Italy. Our results underline the crucial role of the information about the sequencing of events in the analysis of transition to adulthood.

The paper is structured as follows. In Section 2, we introduce some basic notions of the machine learning and data mining approach, which are partially novel to a social science audience. In Section 3, we present the data we use and discuss some of the basic features of the transition to adulthood in Austria and Italy. Section 4 introduces the experimental setup, and it provides a presentation and discussion of the results. Section 5 contains some final remarks.

2 Machine Learning and Data Mining

Machine learning is one of the core research areas in Artificial Intelligence. These days the most prominent research topic within the field is the inductive analysis of databases. Together with statistics and database technology, this area provides the core methodologies for the rapidly developing field of *Knowledge Discovery in Databases*, also known as *Data Mining* [14], which has recently attracted the interest of industry and is considered by many to be one of the fastest-growing commercial application areas for Artificial Intelligence techniques. Machine learning and data mining systems are used for analyzing telecommunications network alarms, supporting medical applications, detecting cellular phone fraud, assisting basketball trainers, controlling elevators, categorizing celestial bodies, and classifying documents on the World-Wide Web, to name only a few applications. A selection of recent applications in machine learning and data mining can be found in [23], and excellent textbooks for the research area are [24, 32]. Within the social sciences, however—and demography is no exception—these tools have not yet received much attention despite the importance of data-oriented research.

In the remainder of this section, we will briefly introduce the classification problem we are dealing with and discuss two common approaches to solving it: the induction of decision trees and inductive rule learning. It can be safely skipped by readers familiar with these techniques.

2.1 Problem Description

The task that has received the most attention in the machine learning literature is the following: given a database of observations (described with a fixed number of measurements x_i, so-called *features* or *attributes*) and a designated attribute y, the *class*, find a mapping f that is able to compute the class value $y = f(x_1, \ldots, x_n)$ from the feature values of new, previously unseen observations. While there are statistical techniques that are able to solve particular instantiations of this problem, machine learning techniques provide a strong focus on the use of categorical, non-numeric attributes and on the immediate interpretability of the result. This, in particular, is one of the main reasons for the increasing popularity of machine learning techniques in both industry and academia.

Temperature	*Outlook*	*Humidity*	*Windy*	*Golf?*
hot	sunny	high	false	no
hot	sunny	high	true	no
hot	overcast	high	false	yes
cool	rain	normal	false	yes
cool	overcast	normal	true	yes
mild	sunny	high	false	no
cool	sunny	normal	false	yes
mild	rain	normal	false	yes
mild	sunny	normal	true	yes
mild	overcast	high	true	yes
hot	overcast	normal	false	yes
mild	rain	high	true	no
cool	rain	normal	true	no
mild	rain	high	false	yes

Table 1: A sample database

Table 1 shows a small sample database, taken from [29]. Given are four attributes—temperature, outlook, humidity, and windy—that measure certain environmental conditions that might be relevant for a person's desire to go out and play golf. The database contains 14 instances of past behavior of this person: weather conditions along with her decision whether she went out to play golf or not. The learning task is to use this *training set* for deriving a model that is able to predict for new weather conditions whether the person is likely to play golf or not. Note that the focus here is *discrimination* and not *characterisation*. When learning discriminative models, one is interested in a minimal set of features that allow one to discriminate objects of one class from objects of another class. This is different from learning characteristic descriptions, which try to describe the commonalities of all objects of a given class. For example, it is easy to discriminate an elephant from all other mammals by referring to its tusks and trunk. A characteristic description would also have to mention size, weight, ears, skin, etc., in short, everything that is needed to explain what an elephant looks like (as opposed to everything that is needed to recognise an elephant).

2.2 Induction of Decision Trees

The induction of decision trees is one of the oldest and most popular techniques for learning discriminatory models [29]. A *decision tree* is a particular type of classification model that is fairly easy to induce and to understand.[1] Figure 1 shows a sample tree which might be induced from the data of Table 1. Classification of a

[1] In the statistical literature (cf., e.g., [6]), decision trees are also known as *classification trees*. Related techniques for predicting numerical class values are known as *regression trees*. Such techniques are also used for predictive purposes in survival analysis. An interesting application of regression trees to demographic data is [10].

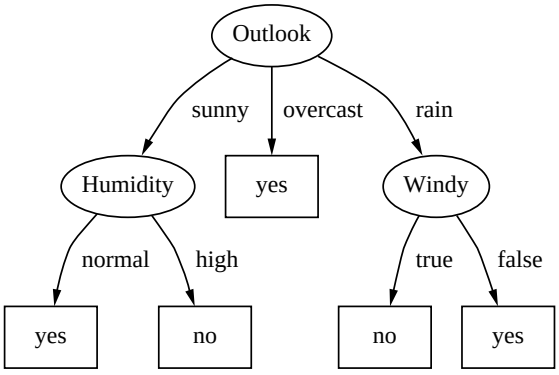

Figure 1: A decision tree describing the dataset shown in Table 1.

new example starts at the top node—the *root*—and the value of the attribute that corresponds to this tree is considered (`Outlook` in the example). Classification then proceeds by moving down the branch that corresponds to the particular value of this attribute, arriving at a new node with a new attribute. This process is repeated until we arrive at a terminal node—a so-called *leaf*—which is not labeled with an attribute but with a prediction. Figure 1 shows leaves as rectangular boxes.

Decision trees are learned in a top-down fashion: the program selects the best attribute for the root of the tree, splits the set of examples into disjoint sets (one for each value of the chosen attribute, containing all training examples that have the corresponding value for this attribute), and adds corresponding nodes and branches to the tree. If there are new sets that contain only examples from the same class, a leaf node is added for each of them and labeled with the respective class. For all other sets, an interior node is added and associated with the best attribute for the corresponding set as described above. Hence, the dataset is successively partitioned into smaller datasets until each set only contains examples of the same class. This condition can always be satisfied unless the training data contains contradictory examples, i.e., examples with the same feature values but different class values.

The crucial step in decision tree induction is the choice of an adequate attribute. To see the importance of this choice, consider a procedure that constructs decision trees simply by picking the next available attribute. The result is a much more complex and less comprehensible tree (Figure 2). Most leaves originate from a single training example, which means that the labels that are attached to the leaves are not very reliable. Although the trees in Figures 1 and 2 will both classify the training data correctly, the former appears to be more trustworthy, and it has a higher chance of correctly predicting the class values of new data.[2] The problem of generating overly complex models that explain the training data but do not generalise well to

[2]This preference for simple models is a heuristic criterion known as *Occam's Razor*, which appears to work fairly well in practice. It is often recalled in the statistical literature on model selection, but it is still the subject of ardent debates within the machine learning community [11].

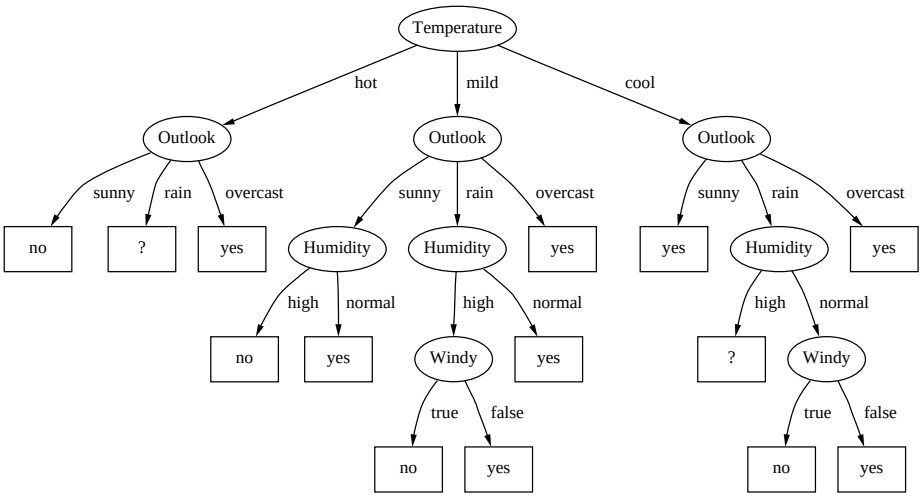

Figure 2: A bad decision tree describing the dataset shown in Table 1.

unseen data is known as *overfitting*.

The attribute selection criterion that is most commonly used in decision tree induction algorithms is *information gain*. It measures the amount of information that can be gained about the class membership of the training examples by splitting the examples using attribute x_i. The formula for choosing the most promising attribute x_r is

$$x_r = \arg\max_i \left[-\sum_c p(c) \log p(c) + \sum_v p(x_i = v) \sum_c p(c|x_i = v) \log p(c|x_i = v) \right] \tag{1}$$

where c iterates over all class values and v iterates over all possible values of the attribute x_i. The first term measures the information content of the class distribution in the current set. From this term, the second term is subtracted (note that the sums are negative!), which measures the weighted average of the information contents of the sets that result from splitting using attribute x_i. Maximising this difference results in the choice of the attribute which reduces the heterogeneity in the class distributions in the successor nodes the most. This enforces a fast convergence towards nodes where the majority of the examples belong to the same class. Of course, this "greedy" procedure will only find locally optimal choices for the attribute. For details we refer the reader to [29, 30, 24, 32].

Obviously, the choice depends on the quality of the estimates for the probabilities used in Formula 1. Typically, these are estimated with the corresponding frequencies in the datasets. However, due to the successive splitting of the data, the datasets that are used for estimating these probabilities become smaller and smaller; hence, the attribute choices become increasingly inaccurate. This effect is strengthened if the data contain *noise*, i.e., erroneous measurements for some

IF Outlook = sunny **AND** Humidity = normal
THEN yes
IF Outlook = sunny **AND** Humidity = high
THEN no
IF Outlook = overcast
THEN yes
IF Outlook = rain **AND** Windy = true
THEN no
IF Outlook = rain **AND** Windy = false
THEN yes

Figure 3: A rule set describing the dataset shown in Table 1

attribute or class values (which is quite common in real-world applications). This is the main reason why state-of-the-art decision tree induction techniques employ a post-processing phase in which the tree generated with the above procedure is simplified by *pruning* branches and nodes near the leaves. In effect, this procedure replaces some of the interior nodes of the tree with a new leaf, thereby removing the subtree that was rooted at this node. The exact details of this procedure are beyond the scope of this paper (we again refer to [30]), but it is important to note that the leaf nodes of the new tree are no longer *pure* nodes, i.e., they no longer contain training examples that all belong to the same class. Typically, this is simply resolved by predicting the most frequent class at a leaf. The class distribution of the training examples within the leaf may be used as a reliability criterion for this prediction.

2.3 Induction of Rule Sets

Another important machine learning technique is the induction of rule sets. Rule sets are typically simpler and more comprehensible than decision trees. To see why, note that a decision tree can also be interpreted as a set of **IF–THEN** rules. Each leaf in the tree corresponds to one rule, where the conclusion of the rule is the label of the leaf, and the conditions encode the path that is taken from the root to this particular leaf. Figure 3 shows the set of rules that corresponds to the tree in Figure 1. Note the rigid structure of these rules. For example, the first condition always uses the same attribute, namely, the one used at the root of the tree. As we shall see, this does not have to be the case if the rules are learned directly.

The main difference between the rules generated by a decision tree and the rules generated by a rule learning algorithm is that the former rule set consists of non-overlapping rules that span the entire instance space (i.e., each possible combination of feature values will be covered by exactly one rule), while the latter consists of potentially overlapping rules that need not span the entire instance space. In this case, mechanisms for tie breaking (i.e., which rule to choose when more than one covers the given example) and default classifications (what classifi-

IF Outlook = overcast
THEN yes
IF Humidity = normal
THEN yes
IF Humidity = high
THEN no
DEFAULT yes

Figure 4: A smaller rule set describing the dataset shown in Table 1

cation to choose when no rule covers the given example) are needed. Typically, one prefers rules with a higher ratio of correctly classified examples from the training set.

Figure 4 shows a particularly simple rule set which uses two different attributes in its first two rules. Note that these two rules are overlapping, i.e., several examples will be covered by more than one rule. For instance, examples 3 and 10 are covered by both the first and the third rule. These conflicts are typically resolved by using the more accurate rule (the first one in our case). Also note that these rule sets make two mistakes (the last two examples). These might be resolved by resorting to a more complex rule set (like the one in Figure 3) but as stated above, it is often more advisable to sacrifice accuracy in the training set for model simplicity to avoid overfitting. Finally, note the default rule at the end of the rule set. This is added for the case when certain regions of the data space are not represented in the training set.

It is beyond the scope of this paper to provide a detailed description of algorithms that learn such rule sets. However, the underlying ideas are quite similar to the ideas used in decision tree induction. The key idea to rule learning is that, instead of successively splitting the example set into regions with increasingly uniform class distributions (in the literature, this strategy is also known as *divide-and-conquer* learning), rule learning algorithms immediately try to focus on regions in which a certain class prevails. This is done by learning a single rule first, removing all examples that are covered by this rule from the training set, and repeating this procedure with the remaining examples (this strategy is also known as *separate-and-conquer* learning). Each rule is learned by picking the best condition (i.e., the test for the presence of a single attribute value that identifies the subset of examples with the most uniform class distribution) and adding it to the rule until the conjunction of all conditions in the rule covers only examples from a single class. Again, pruning is a good idea for rule learning, which means that the rules only need to cover examples that are *mostly* from the same class. It turned out to be advantageous to prune rules immediately after they have been learned, i.e., before successive rules are learned [15]. For a detailed survey of rule learning algorithms we refer the reader to [16].

3 Motivation and data

The transition to adulthood is one of the areas in the sphere of life course events where present-day European countries exhibit a high behavioral heterogeneity [5, 18]. In some countries events are experienced at an early age, while they are postponed to later ages in others. The sequencing of events is also very different, as is sometimes the quantum [8]. These differences, which are linked to cultural and historical patterns, present opportunity structures, and institutional arrangements, are even clearly visible if one considers neighboring countries. In this paper we study Austria and Italy. The choice of these two countries is justified by the different patterns of transition to adulthood they exhibit—this provides us with a clear benchmark, and some prior knowledge, with which we confront the method. In Austria, the duration of education is quite standardised, and the vocational training system allows for a potentially smooth transition from school to work. Further, leaving home occurs to a great extent before marriage, and there is a traditionally high share of births outside of cohabiting (married or unmarried) unions. In Italy, the duration of formal education and entry into the labor market are experienced in a rather heterogeneous way. Leaving home occurs at a late age—the latest age observed among Western countries for which data are available. And leaving home is highly synchronised with marriage. It is not common to leave home before finishing education. Finally, childbearing outside of marriage is still less common that in other European countries[3], respectively.

The data for our analysis originate from the Austrian and Italian Fertility and Family Surveys (FFS), which were conducted between December 1995 and May 1996 in Austria and between December 1995 and January 1996 in Italy. Both surveys were part of a large-scale comparative program co-ordinated by the Economic Commission for Europe of the United Nations. The survey design provided independent samples of men and women in both countries. In Austria, 4,581 women and 1,539 men were interviewed, in Italy 4,824 women and 1,206 men. In Austria respondents were selected from the population aged 20 to 54, while in Italy the age range was 20 to 50. Hence, the Austrian FFS covers birth cohorts from 1941 to 1976, while the Italian FFS only includes cohorts from 1946 to 1976. To avoid differences due to sampling design, we opted to restrict the Austrian data set to the same cohorts as covered in the Italian survey. Furthermore, we excluded records with missing or incorrect values for the timing of events that are included in our analysis. The final dataset contained 11,107 individuals (*examples* in machine learning terms), 5,325 of which were of Austrian and 5,782 of Italian origin.

In the FFS, retrospective histories of partnerships, births, employment, and education (in a more or less complete fashion) were collected on a monthly time scale, which allows us to study the timing, sequencing, and quantum of events in the transition to adulthood. In this study, we analyse the timing and quantum of

[3] Specific descriptions and analyses of the transition to adulthood in Austria and Italy are provided in [27] and in [3].

General Descriptors

sex	female, male
birth cohort (5 years)	1946-50, 1951-55, 1956-60, 1961-65, 1966-70, 1971-75
birth cohort (10 years)	1946-55, 1956-65, 1966-75
age	age at interview in years

Quantum

education finished?	yes, no
had job?	yes, no
left home?	yes, no
formed union?	yes, no
married?	yes, no
had child?	yes, no

Timing

education	age at end of education
first job	age at first job
left home	age at leaving home
union	age at first union
marriage	age at first marriage
children	age at the birth of first child

Ages are measured in years.
If the event has not yet occurred, the interview date is used.

Sequencing

education / job	$<$, $>$, $=$, n.o.
education / left home	$<$, $>$, $=$, n.o.
education / union	$<$, $>$, $=$, n.o.
education / marriage	$<$, $>$, $=$, n.o.
education / children	$<$, $>$, $=$, n.o.
first job / left home	$<$, $>$, $=$, n.o.
first job / union	$<$, $>$, $=$, n.o.
first job / marriage	$<$, $>$, $=$, n.o.
first job / children	$<$, $>$, $=$, n.o.
left home / union	$<$, $>$, $=$, n.o.
left home / marriage	$<$, $>$, $=$, n.o.
left home / children	$<$, $>$, $=$, n.o.
union / marriage	$<$, $>$, $=$, n.o.
union / children	$<$, $>$, $=$, n.o.
marriage / children	$<$, $>$, $=$, n.o.

For each possible combination of timing variables, their relative order is computed, or "n.o." is used if both *events have not yet (i.e. before the interview date) occurred.*

Table 2: Variables used in the experiments

leaving formal education, entering the first job, leaving the parental home, entering first union, entering first marriage, and having a first child, together with their pairwise sequencing. If individuals have not experienced an event, we consider in our analyses variables explicitly indicating that they are censored. Such variables are used as information concerning the quantum of the event.

Two peculiarities of the data need to be mentioned. First, the Austrian FFS only allows one to know when the respondent left home for the last time before the interview, while the Italian FFS explicitly asked when the respondent left home for the first time. This difference should not, however, be a big problem in our comparative analysis: we will be more conservative in comparisons and underestimate differences, as Austrians leave home much earlier than Italians do anyway. Secondly, several problems arise when one wishes to compare educational histories across countries even using FFS comparative surveys [12]. For instance, a considerable number of respondents in Austria (1,639 out of a total of 6,020 respondents) have not indicated any educational level beyond 'Pflichtschule', which is completed at age 15 in Austria. Although education was mandatory until the age of 14 for the Italian cohorts taken into account here (which is already a significant difference from Austria), there are a significant number of people who dropped out before that age. Hence, we should expect institutional and drop-outs differences in the timing of education to show up as an important attribute for differentiating between the Austrian and Italian pathways in the transition to adulthood. We will discuss this further in Section 4.3.

To capture information about timing, sequencing, and quantum, we encoded the information as is shown in Table 2. We used four general descriptors related to sex, age, and birth cohort (with two potentially different categorisations for cohorts). Binary variables are used to indicate whether each of the six events that we employ to characterise the transition to adulthood has occurred in the person's life up to the time of the interview (quantum), similarly to what is done in event history analysis for event/censoring indicators. If an event has occurred, the corresponding timing variable contains the age at which the person experienced the event (computed in years between the birth date and the date at which the event occurred). Finally, to make sequence information accessible to the learning algorithms we performed pairwise comparisons between the dates at which two events occurred. The sequencing relationship, including synchronisation (one date can be smaller or greater than or equal to the other), is encoded as a separate variable.[4] If both events have not occurred, we encode this with a designated value "n.o.". In the case that one of the two events has occurred but not the other, we assume that the one that has occurred occurs earlier, even though the other event might not occur at all in this person's life course. As the time unit for computing sequencing

[4]This approach to making sequence information available to the learner—encoding the additional relations in derived variables—is loosely based on the Linus approach to relational learning (cf., e.g., [19], where learning performance was improved in a medical application by augmenting patient data with additional domain-specific background knowledge that highlighted characteristic combinations of the original measurements).

and synchronisation between two events, we use the month. That is, we use all the information available in the dataset and place a specific emphasis on events that are truly synchronised.[5]

4 Results

4.1 Experimental Setup

We applied decision tree and rule learning algorithms to the dataset described in the previous section in order to detect the key features which distinguish between Austrians and Italians with respect to the timing, sequencing, and quantum of events in the transition to adulthood. We chose the decision tree learning algorithm C4.5 [30] and the rule learning algorithm Ripper [7]. Both algorithms are among the most prominent algorithms in machine learning and are frequently used both for applications and as benchmarks for new algorithmic developments. Their popularity is also due to their wide availability.[6]

In addition to the functionalities described in Section 2, both algorithms are able to handle numerical attributes. C4.5 does this by testing the condition $x_i \leq v$ for each possible value v of the attribute x_i and computing the information that is gained by partitioning the data according to the outcome (*true* or *false*) of this test.[7] This can then be directly compared to the information gain values computed for the categorical attributes. The procedure for Ripper is quite similar. C4.5 also provides an option that allows different values of the same attribute to share the same branch of the tree. We used this option in some of our experiments with the result that the obtained trees were mostly binary.

To estimate the error rates of the learned models we use 10-fold cross-validations [28]. This means that 10 experiments are performed, and in each experiment (each *fold*) a tenth of the data is held out and a model is learned on the remaining nine-tenths. The model learned is then tested on the tenth of data that has been withheld, and the results from these ten (disjoint) test sets are averaged. Note that one cannot simply estimate the error rate of the model from the training data because with increasing model complexity, the algorithms can fit the training data arbitrarily well. However, this fit cannot be expected to hold for new data, a problem that is known as *overfitting*. Our cross-validation folds were *paired* (i.e., the same 10 folds were used for computing the performance estimates of both algorithms, which reduces random fluctuations) and *stratified* (i.e., the number of examples of each class in

[5] Other scholars argue that, since decision-making occurs on a fuzzy time scale, larger intervals for synchronisation should be considered, e.g., a yearly interval [9].

[6] C4.5 can be obtained by buying the companion book [30], or it can be downloaded for research purposes at `http://www.cse.unsw.edu.au/~quinlan/`. C5.0, its commercial successor, is available from `http://www.rulequest.com/`. Ripper is available upon request at `http://www.research.att.com/~diane/ripper.html`.

[7] The algorithms actually implement a more efficient version of this technique, which sorts the values first and tests only values whose successor has a different class value. It can be shown that this procedure produces the same result as testing all values [13].

Feature Set	C4.5		Ripper	
	Error	Size	**Error**	Size
only general descriptors	**45.97**	42.1	**46.83**	15.7
quantum	**33.44**	117.4	**33.99**	38.9
timing	**20.52**	778.2	**19.15**	154.4
sequencing	**17.96**	265.7	**18.40**	48.2
quantum & timing	**19.40**	751.7	**18.62**	188.8
quantum & sequencing	**17.37**	267.0	**17.94**	42.1
timing & sequencing	**15.14**	584.9	**15.18**	116.4
all features	**15.05**	533.7	**14.94**	99.4

Table 3: Error rates (in %) and average size (no. of conditions) for C4.5 and Ripper on different problem representations (estimated by paired 10-fold cross-validations).

each fold was fixed in order to model the class distribution of the original set as closely as possible).

4.2 Quantitative Results with Different Specifications

In order to determine the relative importance of the quantum, timing, and sequencing of events that characterise the transition to adulthood, we performed a series of experiments in which we used different subsets of the available features. Each line of Table 3 shows the achieved performance of one particular feature subset. The first column describes which feature subset is used, the second and third columns show the performance estimates for C4.5 and Ripper, respectively. For both algorithms, we show both the error rate in percentage points and the average size of the model learned (measured by the number of nodes in the decision tree or the number of conditions in the rule set, respectively).[8]

The first line shows the results from using only the four general descriptors shown at the top of Table 2 and none of the quantum, timing or sequencing variables. On this data set, both algorithms achieve error rates that are only slightly better than the error rate of uniformly predicting that all examples belong to the majority class, which has an error rate of 47.94%. These values are included as a benchmark and for checking that the relevant information for satisfactory classifi-

[8] In contrast to the error rate, the size of the models learned could have been measured directly by using the entire set of examples for training. However, we report the average model sizes in the 10 folds of the cross-validation procedure, which came for free as a by-product of the cross-validation procedure. The size of the models learned from the complete training set might be somewhat larger, but their relative order can be expected to be the same.

cation performance is captured by the other variables. The next three lines show the results from using each subset independently, i.e., using only quantum, only timing, or only sequencing information. Among these three, sequencing proves to be most important. Using only sequencing information, both learning algorithms are able to discriminate the life courses of Italians and Austrians with an error rate of about 18%. Quantum information—in the simple encoding that shows only the occurrence or non-occurrence of an event—seems to be the least important. These results are confirmed by the next three lines, which show the performance of each pair of variables. The pair quantum & timing—the only one that does not use sequencing information—produces the worst results, while the timing & sequencing pair, which does not use quantum information, performs best. It should be noted, however, that quantum information is still of importance, as can be seen from the last line, which shows the results obtained using all variables shown in Table 2. Adding quantum to the timing and sequencing information further reduces the error rate (although this decrease is not statistically significant) and also results in simpler models.

In general, the error rates achieved by the two algorithms are about equal. Ripper seems to be a little better in handling timing variables, which may be due to minor differences in the handling of numerical attributes in the two algorithms. But the evidence from this dataset is insufficient to confirm this hypothesis, and there is no systematic comparison along this dimension in the literature. One can say, however, that rule sets are considerably simpler, which also means that they are easier to interpret.

The rule model that uses all features still has about 100 conditions and C4.5's decision tree has more than 500 nodes. Rule sets and trees of that size are very hard to interpret as a whole—one would have to focus on the important aspects (e.g., the areas near the root of the tree). Alternatively, one can make use of the pruning mechanisms of the algorithms to find a reasonable trade-off between simplicity and accuracy.

Both algorithms have a parameter that controls the pruning level of the algorithm. The exact details of the pruning procedures are somewhat different in the two algorithms[9] but both pruning parameters have the purpose of controlling the size of the models learned. This can be used for finding a model size that minimises the estimated error rate as well as for increasing comprehensibility of the models learned.

Figure 5 shows the error and complexity curves for C4.5. It was used with the parameter settings described at the beginning of the next section except for the pruning parameter, which was varied. The measured error rates and model

[9] C4.5 employs *error-based pruning*, which uses a heuristic estimate of the confidence intervals for the accuracy of the class probability estimates at each node [30]. Ripper's pruning is based on the *incremental reduced error pruning technique* [15], which internally splits the available training data into a learning set and a pruning set and uses the latter to fine-tune the rules learned on the learning set.

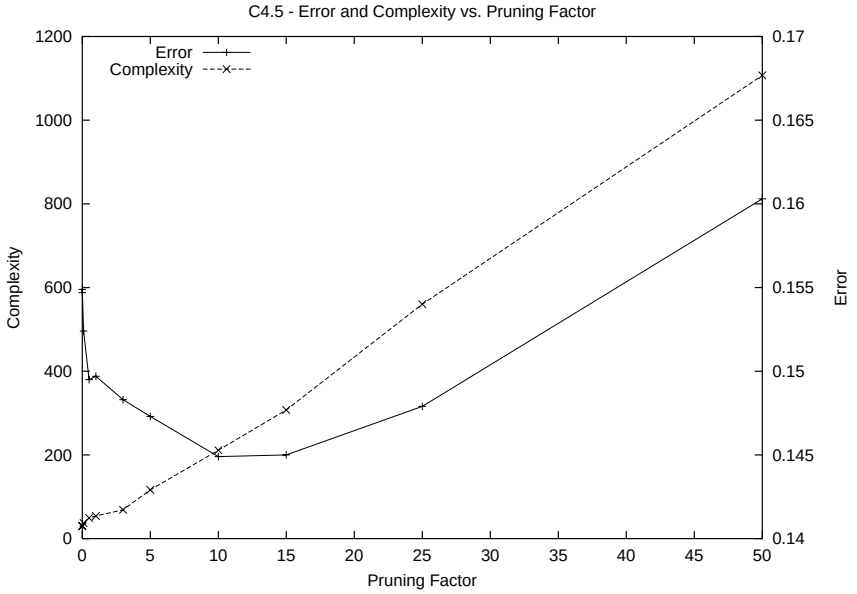

Figure 5: Error rates and model complexity for different settings of C4.5's pruning parameter.

sizes are plotted over the different values of the pruning parameter. The graph clearly exhibits the typical U-shape of error curves: overly complex models may overfit the data while overly simple models will fail to capture some important regularities in the data. Consequently, the best achievable error rate (about 14.45%) is somewhere in the middle, around 10–15 in our case (but this range may be very different for other datasets). C4.5's default value (25), which has been shown to perform reasonably well over a variety of datasets, is close to this. Systematically varying Ripper's pruning parameter led to similar results.

4.3 Interpretation of Results

In this section we analyze the results of two models, one decision tree and one rule set. These models were obtained by using parameter settings that produce more comprehensible models, which are not necessarily optimal. In particular, we allowed C4.5 to combine branches with different attribute values (in default mode it will generate one branch for each possible outcome of a test, i.e., for each possible attribute value), and we specified that Ripper has to learn a rule set for each class (not only for the minor class, which classifies everything else by a default rule). We chose rather aggressive settings for the pruning parameter (0.001 for C4.5), which do not produce models with maximal accuracy, as can be seen from a quick inspection of the graphs in Figure 5. It should be noted, however, that the differences in accuracy are not that large and, more importantly, that the more accurate, complex

models differ from the less accurate, simpler models only in the lower parts of a tree. The upper part, i.e., the choice of the most significant variables, is the same. It is not quite as simple for rule sets, but the starting conditions of the rules are typically more significant and are identical in simpler and more complex models.

Figure 6 presents a simplified tree that uses only 32 nodes. Its estimated error rate is about 15.5%. Its most important attribute is the sequencing of union formation and marriage. Following the right branch that originates at this node and summing over all entries in each leaf implies that 5,445 Italians versus 2,506 Austrians are covered by this path. From these results we can already deduce a first proposition:

> *An important characteristic of the transition to adulthood that identifies Italians is the fact that union formation and marriage are more likely to be synchronised. That is, Italians are much more prone than Austrians to marry directly rather than to start living together before marriage.*

This result is in accordance with the literature stressing the low diffusion of consensual unions in Italy.

In the case when there is no synchronisation of marriage and union formation (i.e. if we follow the left branch that originates at the first node), the age at which education is finished (i.e., the timing of education) is chosen as the next most important attribute. Those who finished their education after the age of 14 are then most likely Austrians (2,807 Austrians vs. 253 Italians). In the case that education is finished before the age of 14, a further attribute is added that compares the sequencing and quantum of education and leaving home. However, the total number of cases captured by this branch is negligible (84 Italians and 12 Austrians), and it might in part be due to the differences in the educational system of Austria and Italy (see Section 3). We will return to this issue later in this section.

If we follow the right branch of the decision tree, that is, in the case either of synchronisation of marriage and union formation or of no experience of these events, the discrimination rules are not as straightforward. However, adding the birth of the first child as the third event and comparing the sequencing and quantum of the date of union formation and the birth of the first child essentially helps to distinguish between Italians and Austrians. Following the right hand branch starting at the second node we are led to a second proposition:

> *If the date of union formation and marriage coincides (or if neither event has yet occurred) Austrians are more likely than Italians to have had a child before this union.*

Though the timing of education is added as a further attribute to this branch, the number of cases included in the final leaf that identifies Italians is negligible.

The classification becomes more complicated if neither event (union formation nor birth of first child) has occurred or if union formation precedes the birth

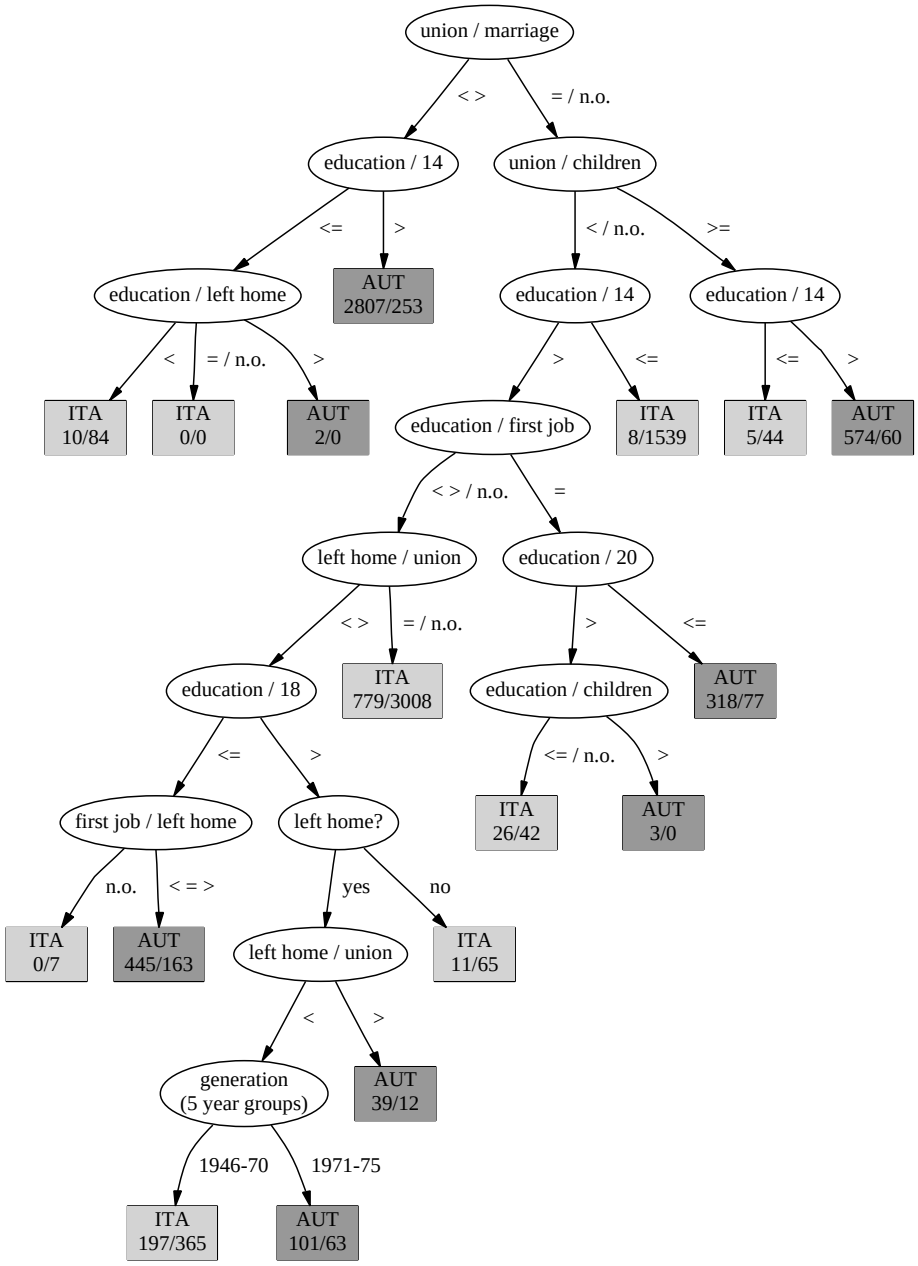

Figure 6: A decision tree applied to data on the transition to adulthood in Austria and Italy. The numbers indicated in each leaf are always ordered such that the first number refers to Austrians and the second number refers to Italians that are covered by this rule.

of the first child. This branch includes 5,341 Italians and 1,927 Austrians. Now the timing of education becomes truly important (as we expected) for further discriminating between Austrians and Italians. Those who completed their education before the age of 15 are most likely to be Italians (1,539 Italians vs. 8 Austrians). Nevertheless, 3,802 Italians and 1,919 Austrians are not yet distinguished in the case that education is completed after the age of 15. A fourth attribute, which refers to sequencing and quantum of the end of education and start of the first job, adds further information. Those for whom the end of education and the start of the first job are synchronised are most likely to be Austrians (119 Italians vs. 347 Austrians). In the case that the sequencing between end of education and start of the first job is indeterminate, or if neither event has yet occurred, a fifth attribute is added. It is at this stage that the age at the time of leaving home is considered. Those for whom leaving home is synchronised with union formation, or for whom neither event has yet occurred, are most likely to be Italians (3,008 Italians vs. 779 Austrians). If these two events are not synchronised (or their sequencing is not yet known), a more complicated decision rule is proposed, which includes the timing of education once again. It also adds the sequencing and quantum of leaving home, union formation, and the start of the first job. Finally, it includes information about cohort membership at the final node. These results lead us to suggest a third proposition:

> *In the case of more traditional patterns in the transition to adulthood, where union formation is synchronised with marriage (or neither of these two events has yet occurred) and the birth of the first child—if it has yet occurred—comes after union formation, the length of education and the sequencing of the end of education and the first job are two further important discriminating factors.*

In particular, those who finished education before the age of 15 are most likely to be Italians. Among those who finished education after the age of 15, Austrians are discerned from Italians by the fact that the end of education and the first job are synchronised or that leaving home and union formation are synchronised (or have not yet been experienced).

As already indicated in Section 3, the timing of education in Austria and Italy reflects institutional differences in the educational system. To test the importance of such institutional differences (as opposed to less formalised differences in the transition to adulthood) we applied the decision tree algorithm to the same data set except that we excluded the length of education as an attribute. The resulting decision tree is presented in Appendix, Figure 8. Its error rate is 2.2% higher than that of the tree in Figure 6, which is allowed to use length of education. A comparison of both trees supports our first two propositions, which were already derived from Figure 6. The sequencing and quantum of the events union formation, marriage, and birth of first child are major attributes for distinguishing between Austrians and Italians, as is the synchronisation (or joint absence) of leaving home

IF home = marriage
THEN *ITA* (592/2851)
IF home = n.o. **AND** union = n.o. **AND** child = n.o.
THEN *ITA* (465/1692)
IF union = marriage **AND** education ≤ 14
THEN *ITA* (9/1308)
IF education ≥ 22 **AND** union = marriage **AND** union ≥ 24
THEN *ITA* (64/541)

IF home $<$ marriage
THEN *AUT* (3476/976)
IF home $>$ marriage **AND** home = yes **AND** education ≥ 15
THEN *AUT* (533/46)
IF education ≥ 15 **AND** job ≤ 18 **AND** education ≤ 18 **AND** union ≤ 21
THEN *AUT* (1468/215)

DEFAULT *AUT* (197/105)

Figure 7: A rule set describing the data on the transition to adulthood in Austria and Italy. The numbers indicated at the end of each rule are always ordered such that the first number refers to Austrians and the second to Italians covered by the rule.

and first union. If we follow the first two right-most branches in the decision tree in Figure 8, the timing of events is added as a further attribute. Obviously, the length of education is now replaced by the age at the start of the first job as the most important attribute. Both events are expected to be closely connected in any life course in which the labor market is entered at all. In comparison to the educational system, however, job histories are less strictly regulated by institutional settings.

To obtain a more compact representation of rules that best discriminate between Austrians and Italians we apply the rule learning algorithm Ripper to the same dataset. Again, we used a setting that favored simplicity over accuracy in order to optimise comprehensibility. The resulting rule set is shown in Figure 7. Its error rate is about 16.5%. One of the rules is the sequencing and synchronisation between leaving home and marriage. They are synchronised for 2,851 Italians as compared to 592 Austrians, while leaving home precedes marriage for 3,476 Austrians versus 976 Italians. The fact that the rule algorithm chooses leaving home and marriage as opposed to union and marriage (which was chosen as the most important attribute in the decision tree algorithm) is not a contradiction. To understand it, we need to recall how both algorithms are designed.

The decision tree chooses the attribute that best discriminates between Austrians and Italians for each representation of the attribute.[10] The attribute union/marriage

[10]Technically speaking, the algorithm determines the performance of each attribute as the average performance across all representations of the attribute and then chooses the attribute that has the best

was chosen because it discriminates best between Austrians and Italians independently of whether both events are synchronised or not or neither event has taken place. On the other hand, the rule learning algorithm takes single representations of attributes (compared to average performance across all representations of one attribute) as the criterion to decide on the best rule. Consequently, the result that Austrians are best distinguished from Italians by their leaving home before marriage does not imply that leaving home after marriage is as good a criterion for discriminating between the two societies.

Besides the important role of synchronisation of events like leaving home, marriage, and union formation, Italians also differ from Austrians in that more of them have not yet (i.e. until the interview date) experienced either of these events. As the second rule states, those who have not yet left home, not yet started a union, and not yet had a child are more likely to be Italians. Related to this pattern, Italians have often be termed as being the 'latest late' as regards events that characterise the transition to adulthood. Austrians experience many of these events at much earlier ages, as is best represented by the last rule.

For the sake of completeness we have added the results of the rule-based algorithm as applied to the data set where we exclude the timing of education as an attribute (see Appendix A, Figure 9).

Though the models learned by the decision tree algorithm differ slightly from those learned by the rule learning algorithm, the results of both also support the proposition that *the sequencing of events in the transition to adulthood is more important than the timing and quantum of these events for discriminating between Austrians and Italians.*

5 Discussion

In this paper we applied techniques developed in machine learning for the analysis of life course data from a comparative perspective. These techniques allow for a high degree of flexibility in the use of data and for problem-specific representations of the available information.

For example, in the transition to adulthood, the key role of the timing, sequencing, and quantum of events such as leaving home, union formation, marriage, birth of the first child, completion of education, and start of the first job is acknowledged from a theoretical point of view. To distinguish between the timing, sequencing, and quantum of events we proposed a novel representation of life course data that captures this information. We then applied machine learning techniques to event data about the transition to adulthood from Austrian and Italian Fertility and Family Surveys. More specifically, we built decision trees and rule sets that aimed to discriminate between Austrian and Italians in terms of characteristics that pertain to the transition to adulthood.

---

average performance across all its representations.

Our main theoretical result is that we have established the key role of the sequencing of events for distinguishing between Austrian and Italian life courses in the transition to adulthood. Information on the timing of events could be regarded as the next best group of features, while information on the quantum of events turned out to be the least important feature set. In terms of the sequencing of events, our findings showed that the synchronisation between events such as leaving home and first marriage and first marriage and union formation is the most important feature for distinguishing the transitional paths of Italian young people from those of Austrians. Information on the timing of events was only of importance as it regards the length of formal education and the age at the start of the first job and only in the case that the algorithm needed to distinguish between Austrians and Italians along more traditional life course patterns. Information on the quantum of events (i.e., whether an event has occurred) was mostly confined to highlighting a well-known characteristic of Italian patterns in the transition to adulthood, namely, the fact that Italians are often the 'latest late' as regards events such as leaving home, union formation, marriage, and birth of the first child. However, we should stress that by focusing on the transition to adulthood and on non-repeatable events, we have unavoidably underscored the importance of the quantum of events. We think that in other applications the quantum may play a major role.

The adoption of the machine learning approach we proposed allows one to look at life courses from a holistic perspective. This perspective becomes even more important in the case of comparative studies, i.e., if one tries to differentiate between two groups of individuals. Moreover, the results of decision trees and rule sets provide a non-technical audience with a clear representation of results. This is not possible with the techniques currently in use, either because they do not start from a holistic perspective or because they do not give results which are clearly interpretable. We thus foresee many applications for the techniques discussed here in life course research and more generally in demography and sociology. In general, such techniques can be applied to various kinds of comparative research. Social dynamics arising from the comparison of cohorts, and gender comparisons within a society are also envisageable as potential research areas. The public availability of software for these techniques is definitely a great advantage. Finally, the representation of life course events in terms of timing, sequencing, and quantum we have adopted can be regarded as a further novelty to come out of this paper. This representation can possibly also be adopted in analyses using different techniques based on other statistical models.

Disclaimer

The views expressed in this paper are the authors' own views and do not necessarily represent those of the Max Planck Institute for Demographic Research or the Austrian Research Institute for Artificial Intelligence.

Acknowledgements

The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry of Education, Science and Culture. The experimental work in this paper was greatly facilitated by a set of tools developed within the ESPRIT long-term research project METAL (project nr. 26357). We wish to thank their author, Johann Petrak, for invaluable help in using these tools.

The authors wish to thank Karl Brehmer for editorial comments, and the Advisory Group of the FFS programme of comparative research for its permission, granted under identification number 75, to use the FFS data on which this study is based.

References

[1] Abbott A., 1995, Sequence Analysis: New Methods for Old Ideas, *Annual Review of Sociology*, 21: 93-113.

[2] Abbott A., Tsay A., 2000, Sequence Analysis and Optimal Matching Methods in Sociology. Review and Prospect, *Sociological Methods & Research*, 29(1): 3-33.

[3] Billari F.C., 2000, *L'analisi delle biografie e la transizione allo stato adulto. Aspetti metodologici e applicazioni ai dati della Seconda Indagine sulla Fecondità in Italia*, Cleup Editrice, Padova.

[4] Billari F.C., Piccarreta R., 2000, Studying demographic life courses with sequence analysis, *paper under review*.

[5] Billari F.C., Philipov D., Baizán Muñoz P., 2000, Leaving home in Europe. An overview on cohorts born around 1960, *Paper prepared for the Workshop on Leaving Home-A European Focus, Max Planck Institute for Demographic Research, Rostock, 6–8 September*.

[6] Breiman L., Friedman J., Olshen R., Stone C., 1984, *Classification and Regression Trees*, Wadsworth & Brooks, Pacific Grove, CA.

[7] Cohen W.W., 1995, Fast effective rule induction, in A. Prieditis and S. Russell, editors, *Proceedings of the 12th International Conference on Machine Learning (ML-95)*, pp. 115–123, Lake Tahoe, CA. Morgan Kaufmann.

[8] Corijn M., 1999, Transitions to adulthood in Europe for the 1950s and 1950s cohorts, *CBGS-Werkdocument*, 4, Brussels.

[9] Courgeau D., Lelièvre É., 1992, *Event History Analysis in Demography*, Clarendon Press, Oxford.

[10] De Rose A., Pallara A., 1997, Survival Trees: An Alternative Non-Parametric Multivariate Technique for Life History Analysis, *European Journal of Population*, 13: 223-241.

[11] Domingos P., 1999, The role of Occam's Razor in knowledge discovery, *Data Mining and Knowledge Discovery*, 3(4): 409–425.

[12] Dourleijn E., Liefbroer A.C., Beets G.C.N., 2000, The measurement of educational attainment in the FFS: Comparing the ISCED-classification with information from educational histories in 17 European countries, *Paper prepared for the FFS Flagship Conference, Brussels, May 29-31*.

[13] Fayyad U.M., Irani K.B., 1992, On the handling of continuous-valued attributes in decision tree induction, *Machine Learning*, 8: 87–102, Technical Note.

[14] Fayyad U.M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. (eds.), 1995, *Advances in Knowledge Discovery and Data Mining*, AAAI Press, Menlo Park.

[15] Fürnkranz J., 1997, Pruning algorithms for rule learning, *Machine Learning*, 27(2): 139–171.

[16] Fürnkranz J., 1999, Separate-and-conquer rule learning, *Artificial Intelligence Review*, 13(1): 3–54.

[17] Hogan D.P., 1978, The variable order of events in the life course, *American Sociological Review*, 43: 573-86.

[18] Kiernan K., 1999, Childbearing outside marriage in Western Europe, *Population Trends*, 98: 11-20.

[19] Lavrač N., Džeroski S., Pirnat V., Križman, V., 1993, The utility of background knowledge in learning medical diagnostic rules, *Applied Artificial Intelligence*, 7: 273–293.

[20] Liefbroer A.C., 1999, From Youth to Adulthood: Understanding Changing Patterns of Family Fromation from a Life Course Perspective, in [31]

[21] Lillard L.A., 1993, Simultaneous equations for hazards. Marriage duration and fertility timing, *Journal of Econometrics*, 56: 189-217.

[22] Marini M.M., 1987, Measuring the Process of Role Change during the Transition to Adulthood, *Social Science Research*, 16: 1-38.

[23] Michalski R.S., Bratko I., Kubat M. (eds.), 1998, *Machine Learning and Data Mining: Methods and Applications*, John Wiley & Sons.

[24] Mitchell T.M., 1997, *Machine Learning*, McGraw Hill.

[25] Modell J., Furstenberg F.F. Jr., Hershberg T., 1976, Social Change and Transitions to Adulthood in Historical Perspective, *Journal of Marriage and the Family*, 38: 7-32.

[26] Mulder C., Wagner M., 1993, Migration and Marriage in the Life Course: a Method for Studying Synchronized Events, *European Journal of Population*, 9(1): 55-76.

[27] Nowak V., Pfeiffer C., 1998, Transition into Adulthood, *Working Paper 8*, Austrian Institute for Family Studies.

[28] Stone M., 1974, Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society B*, 36: 111–147.

[29] Quinlan J.R., 1986, Induction of decision trees, *Machine Learning*, 1: 81–106.

[30] Quinlan J.R., 1993, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA.

[31] van Wissen L.J.G., Dykstra P.A. (eds.), 1999, *Population Issues: An Interdisciplinary Focus*, Kluwer Academic/Plenum Publishers, New York.

[32] Witten I.H., Frank E., 2000, *Data Mining — Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Publishers.

Appendix

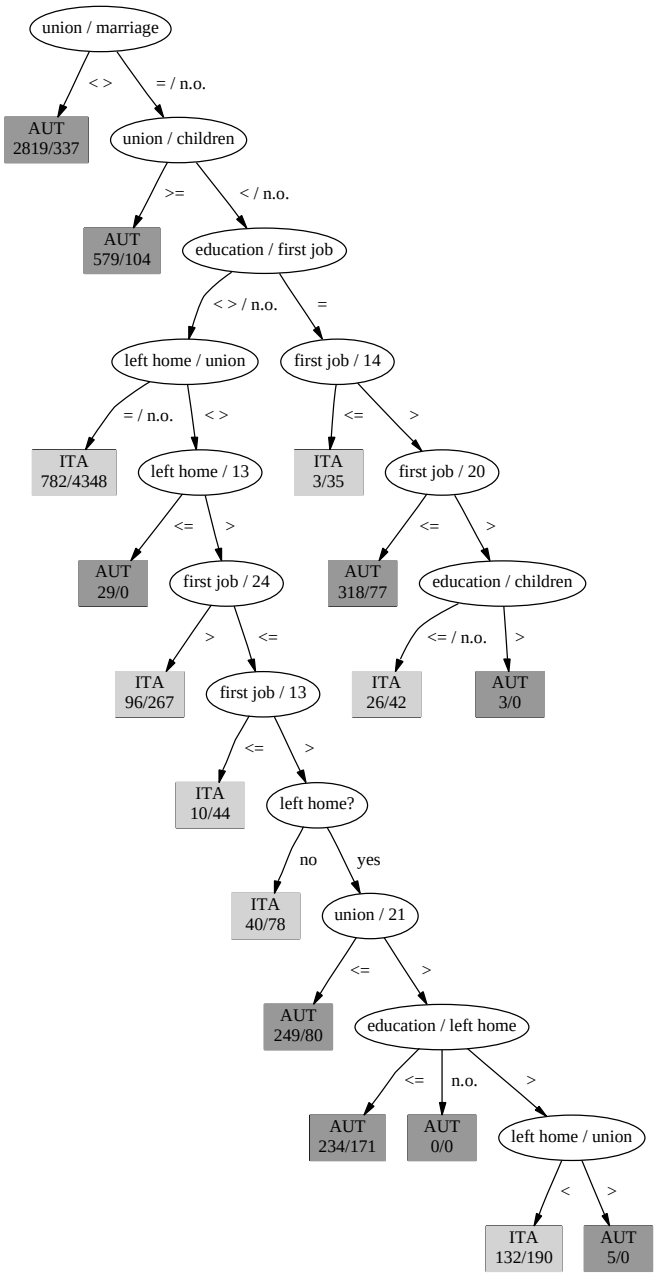

Figure 8: A decision tree applied to data on the transition to adulthood in Austria and Italy excluding the timing of education as an attribute. The numbers indicated in each leaf are always ordered such that the first number refers to Austrians and the second to Italians covered by this rule.

IF home $=$ marriage
THEN *ITA* (592/2851)
IF home $=$ n.o. **AND** child $=$ n.o.
THEN *ITA* (529/1728)
IF union $=$ marriage **AND** job ≥ 22
THEN *ITA* (215/1419)

IF home $<$ marriage
THEN *AUT* (3476/976)
IF home $>$ marriage **AND** home $=$ yes
THEN *AUT* (534/80)
IF job < 20 **AND** education $=$ job **AND** job ≥ 17
THEN *AUT* (824/57)
IF home $>$ union **AND** union $<$ marriage
THEN *AUT* (458/45)

DEFAULT *AUT* (99/79)

Figure 9: A rule set describing the data on the transition to adulthood in Austria and Italy excluding the timing of education as an attribute. The numbers indicated at the end of each rule are always ordered such that the first number refers to Austrians and the second to Italians covered by this rule.