

Sequence Analysis in Demographic Research

Francesco C. Billari

Max Planck Institute for Demographic Research
Rostock, Germany

Abstract

This paper examines the salient features of sequence analysis in demographic research. The new approach allows a holistic perspective on life course analysis, and is based on a representation of lives as sequences of states. Some of the methods for analysing such data are sketched, from complex description to optimal matching to monothetic divisive algorithms. After a short illustration of a demographically relevant example, the needs in terms of data collection and the opportunities of applying the same approach to synthetic data are discussed.

Résumé

On examine ici les principaux éléments de l'analyse par séquence en démographie. Cette nouvelle technique permet une perspective unifiée de l'analyse du cours de la vie, en représentant la vie comme une série d'états. Certaines des méthodes pour de telles analyses sont décrites, en commençant par la description complexe, pour considérer ensuite les alignements optimales, et les algorithmes de division. Après un court exemple en démographie, on considère les besoins en données et les possibilités d'application aux données synthétique.

Key words: Sequence analysis, life course, longitudinal data, synthetic biographies, monothetic divisive algorithm.

Holistic Perspectives on Life Course Data Analysis: The Strong vs. the Pragmatic View

The study of life courses has in recent years focused increasingly on event histories. This is naturally connected to the way life events are seen as the ultimate *explanandum* by the life course research stream (Mayer and Tuma, 1990; Giele and Elder, 1998), as well as to the productiveness of this approach. In fact, with the help of event history analysis it is possible to study the intersection and mutual interdependencies between parallel careers of individuals and between potentially interdependent individuals such as couples. In addition, one can analyse the impact of variables situated at different levels of aggregation on individual behaviour (Blossfeld and Rohwer, 1995). Event history analysis has gradually become important in demographic research. For instance, a substantial part of the design of the recent Fertility and Family Survey Project co-ordinated by the United Nations Economic Commission for Europe was influenced by the wish to use event history modelling techniques in the analyses.

Though it has played a crucial role, event history analysis cannot address every interesting issue that arises in life course research¹. Event history analysis, as its name suggests, focuses on the time to occurrence (or non-occurrence) of specific events in the life course (Yamaguchi, 1991). By looking at its (latent) dependent variable, the transition or hazard rate, event history analysis expresses well the idea of what we may call an ‘event-oriented’ approach in life course research. Nevertheless, by focusing on time-to-event, researchers may miss a general overview of life courses, thus failing to adopt a ‘holistic’ perspective that sees life courses as one meaningful conceptual unit. A holistic perspective calls for efforts to analyse life courses in their wholeness, instead of specific events or combinations of events, as dependent variables. Taking this as a starting point, I shall focus in this paper on the idea of coupling the widespread event-oriented perspective with a holistic perspective on life courses. I shall review and discuss an approach that may prove useful for this purpose; in fact, it has already been proved useful.

Different theoretical approaches share a common goal: to analyse the life courses of individuals starting from a certain point in time and, where applicable, up to a certain point in time. But why should one be interested in choosing, in addition to an event-oriented approach, a holistic perspective on life courses? I see at least two theoretically significant ways of justifying this interest, each with a distinct emphasis on the strength of underlying hypotheses on individual behaviour. For simplicity, I shall call them the “strong” and the ‘pragmatic’ views.

The ‘strong’ viewpoint, which is present for instance in some of the traditional literature in neo-classical economics, sees life courses as subject to accurate inter-temporal planning and hence as outcomes of utility maximisation (see, for

instance, the review of Deaton and Muellbauer, 1980,Ch.12). The classical Modigliani's life-cycle hypothesis, Friedman's permanent income concept, and the ensuing stream of research are good examples. Such a view has also led to empirical study of long-term plans in life courses and their consistency, as well as to critiques of a dynamic programming view of lives from experimental economists (see the literature review of Camerer, 1995). In other words, according to the 'strong' view, life courses are mainly the outcomes of planning (in an uncertain world with possibly external stochastic shocks). A holistic perspective is thus hypothesised to be present in the behaviour of individuals themselves, directly embedded in the behavioural assumptions.

In psychology too, individual life courses are considered emerging from internalised timetables (Heckhausen, 1999). In the sociodemographic literature, the notion of strategy, or 'life-planning', has been emphasised. As Giddens says: "In a world of alternative lifestyle options, strategic *life-planning* becomes of special importance. (...) Life-planning is a means of preparing a course of future actions mobilised in terms of the self's biography" (1991, p. 85). Settersten and Mayer (1997) maintain that the older concepts of the 'life course' were holistic, with more or less explicit reference to biological structures. This is no longer true in the life course literature, which also started out with a strong emphasis on internalised life calendars and schedules. So, for a theoretical approach that assumes the individuals look holistically at their own lives, it is undoubtedly necessary to have tools that allow us to follow the same perspective, and to treat the life course as a conceptual unit.

According to the 'pragmatic' viewpoint, the life course as a conceptual unit is thought of as being a contingent result of sequence of events. Following this viewpoint, researchers focus principally on events when they wish to explain individual behaviour. This view (see, e.g., Rohwer, 1994 and Blossfeld and Rohwer, 1995) is justified by its links to the philosophical notions of causality and time. Even if one takes such a position, a holistic perspective is useful as a way to describe and to summarise the past history of individuals. One can also play the role of an historian. In life course research, it is fundamental to study the timing of events, their sequencing, the duration of time spent in states, and the spacing between events (Settersten and Mayer, 1997). Comparative research across countries or regions or cohorts is one of the examples where the life courses as a whole conceptual unit might provide particular insights.

If we accept that studying life courses in their wholeness is worthwhile and important, it is then necessary to complement the event-oriented techniques with those that can analyse them as a unit. In this paper, I deal with the *sequence analysis* approach to this topic, introduced to the social sciences by Abbott. In Section 2, I give a brief introduction to the sequence representation of life courses. Then, I review some of the approaches to analysing life courses represented as sequences. In Section 4, I briefly describe a demographic application of this approach. The issue of data collection is dealt with in section

5. Section 6 discusses the use of sequence analysis as a tool for analysing synthetic biographies. Finally, I consider some of the open questions and opportunities. In the appendix the reader will find a concise reference to existing software on sequence analysis.

Life Courses in a “Word” : The Sequence Representation

Different approaches² have been proposed in the literature for studying life courses as whole conceptual units from a quantitative point of view. Nowadays, at least in sociology, there is an increasing agreement about focusing on the set of techniques known as *sequence analysis* (for a review, see Abbott and Tsay, 2000; Abbott, 1995). The basic idea is to represent each life course or trajectory in the life course as a ‘word’ or, to be precise, as a string of characters (also numerical). This representation is thus identical to the one used to code DNA molecules (Waterman, 1995). By and large, one focuses on a time window with a precise beginning and endpoint (two specific ages). Then, the technical question arising is how to analyse such strings in a meaningful and easily interpretable way.

Abbott (1995) makes a distinction between ‘non-recurrent sequences’ those where a character may not repeat at all, and ‘recurrent sequences,’ where repetition is possible, exactly as in molecular biology. As a simple example, think of two characters: A and B. These two characters can give rise to five non-recurrent sequences: an empty sequence, A, B, AB and BA. There is an infinite number of possible sequences if the characters may be repeated. For our purposes, the distinction between recurrent and non-recurrent sequences can be extended to the concepts of ‘sequences of events’ and ‘sequences of states’ (Billari and Piccarreta, 2000). We shall make very brief reference to sequences of events and then discuss sequences of states in greater detail.

When representing a life course as a sequence of events, one normally assigns a letter (or a number) to an event, and the ordering of events gives the ordering of letters in the word. Let us, for example, represent the union behaviour of an individual. He/she first forms a cohabiting union (event denoted by C), then gets married (M), then gets divorced (D) and remarries. A representation of this life course via a sequence of events would be: CMDM. The main advantages of such a representation are simplicity and compactness. The main problem with it is that one cannot explicitly take into account the time between events; in fact, this approach makes use of time implicitly in the sense that events are ordered. In particular, it is not possible to represent events that occur simultaneously, which can sometimes be the case when analysing demographic behaviour. The representation is, however, interesting when either the number of events is low or the complexity of life courses is limited, in the sense that they are concentrated in patterns that have exactly the same representation.

Reading Behind the Words: Analysing the Life Course as a Sequence of States

What analytical strategies can we follow when we have access to a set of sequences of states? Standard distribution-based methods will not work simply as a consequence of the complexity of the problem. With sequences on a monthly time scale and with a long enough time span (e.g. 20 years) on a yearly time scale, the probability that two sample members can be represented by the same sequence becomes very low, tending towards zero. We therefore need techniques tailored to the problem. In addition, it is not a straightforward step to directly employ the methods used in computational biology, because the problems of the two fields are clearly of a different nature. Biological sequences are typically very long, and each of them has complex internal patterns with a vast number of states and state changes and a specific meaning (e.g., see Myers, 1995). In the social sciences, it is normal to have a large number of sequences of relatively short length (compared to biological sequences), and we are hardly interested in each individual sequence. The aim is rather to discover regularities in behaviour of a group of individuals. Here, I shall briefly review some of the approaches proposed in the social science literature.

Description Based on Features of Individual Sequences

While the description of individual life courses may be effective, especially with graphical tools (see, e.g., the descriptions provided by BioBrowser, Statistics Canada, 1999³ and the monograph by Wehner, 1999), it is difficult to represent more than a handful of life courses. So, we need some other tools. Following Rohwer and Trappe (1999), we may distinguish two different ways of describing sequences of states, one 'cross-sectional' and the other 'longitudinal.' The *cross-sectional approach* to sequence description takes as its starting point an origin on the time scale that is common to all sequences (say, a specific age like 15 for fertility analysis)⁴. The purpose is to synthesise, in various ways, the state distribution at that point in time.

The *longitudinal approach* is based on the idea that individuals cannot be simply classified according to their characteristics observed at only one point in time. So, observation time and past life history should be considered as being meaningful when describing sequence data. The purpose is then to synthesise the past history of individuals up to a certain point in time. The longitudinal approach can also be seen in a dynamic way: the synthesis we construct evolves across time.

While I cannot give here a full account of the different paths that can be followed (see Rohwer and Pötter, 2000), I wish to mention here an interesting idea commonly used with biosequences, namely the search for meaningful patterns. To give an example, let us go back to the single-cohabitation-marriage

sequence illustrated above. It would be interesting, for instance, to know how often first marriages are preceded by unmarried cohabitation. We can indicate this as a pattern in a sequence: S..SC..CM, where ‘..’ means permanence in a given state. Another example of the idea would be the pattern: S*C*M, where * stands for an arbitrary (and possibly empty) sequence of states. The occurrence of patterns in different sequences or within individual sequences might provide insights for the analyses we are interested in. Obviously, the search for patterns can also be seen from a dynamic perspective, namely the presence of a pattern as a function of time.

Computer graphics, optimally with colour capability, are particularly useful in the description of sequences of states. We refer the reader to Rohwer and Trappe (1999), Rohwer and Pötter (2000) as well as Billari and Piccarreta (2000) for a comprehensive discussion of this issue.

Optimal Matching Analysis

The ‘optimal matching analysis’ is based on the notion of similarity or dissimilarity between pairs of sequences. This method has been used for the alignment of biosequences. The initial question is the following: what would it mean to say that two sequences (life courses) are more similar than two other sequences? This is, of course, a complicated question, one that cannot be answered easily. We may well come to the conclusion that the complexity of whole life courses does not allow for comparison in terms of a single (one-dimensional) metric. We shall, however, discuss optimal matching as one possible way of analysing sequences.

The basic idea behind optimal matching is to measure the dissimilarity of two sequences by considering how much effort is required to transform one sequence into the other one. Transforming sequences entails three basic operations in its most elementary method:

- 1) *insertion*: a state is inserted into the sequence;
- 2) *deletion*: a state is deleted from the sequence;
- 3) *substitution*: a state is substituted by another one.

To each elementary operation a specific cost can be assigned, and the cost of applying a series of elementary operations can be computed as the sum of the costs of single operations. The distance between two sequences can thus be defined as the minimum cost of transforming one sequence into the other one. For example, if insertions and deletions cost one unit and substitutions two units, the cost of transforming the sequence SSCMMM into the sequence SCCMM is 2 units. Specific dynamic programming algorithms assure that the minimum cost is effectively sought out (Sankoff and Kruskal, 1983; Waterman, 1995).

The computed distance thus takes into account the entire sequence and not just present states. As a result, one obtains a distance matrix. This can be employed as an input for any kind of analysis requiring proximity data (e.g., clustering and multidimensional scaling). As a result of a series of seminal papers by Abbott, most of the sociological literature makes use of this method. Chan (1999) gives a compact and useful review of the application of optimal matching analysis in life course research. Wu (2000) gives a critical view on this approach.

Unfortunately, this approach also has some drawbacks. First, we must take each sequence as a whole and view life courses as dynamically evolving through time. One possible way to cope with this problem is to do a series of sequence analysis for each time period in the observational window. This would then result in a sequence of distance matrices. The second drawback is that it can be difficult to understand which variables in the definition of specific clusters are more relevant than the others. As Halpin and Chan (1998) state, “while the clusters are very easy to characterise in a general way, it is impossible to characterise them formally and exhaustively, that is, to define rules which will replicate the clusters exactly or close to exactly.”

Clustering Binary Sequences

A proposal by Billari and Piccarreta (2001) overcomes this last-mentioned problem of the optimal matching method by explicitly building meaningful groups and by using algorithms for the clustering of binary variables. The algorithm applies to a series of parallel sequences that can be represented by binary variables (such as in the example with births discussed in an earlier Section). For a meaningful interpretation of the algorithm it is necessary that the events be non-renewable. For example, a value of 1 at a certain point in time must imply a 1 at all the following points in time. Billari and Piccarreta's proposal is to use a monothetic divisive clustering algorithm. The algorithm is hierarchically divisive, in the sense that the entire sample is first divided into two groups, and each group is further split into two subgroups. The procedure is iterated until each individual belongs to his/her own group. The algorithm is also monothetic: each group is divided into two subgroups according to the values of a single binary variable. To perform the splitting, it is also necessary to select a single relevant variable in such a way that the two subgroups are characterised by the maximum homogeneity within and the maximum heterogeneity between. Heterogeneity can be measured using either Gini Index or any entropy measure. The splitting procedure can also be represented by means of a tree diagram.

The main advantage of this algorithm is that it leads to easily interpretable clusters: the groups obtained are in fact perfectly characterised by the presence (or absence) of certain attributes (those implied by the splitting variables). Another interesting feature of the algorithm lies in the fact that enables us to identify the most relevant variables in the clustering process and to rank these

variables according to their importance in the clustering process. One will in fact attribute a higher discriminating power to variables that induce the first splits. A major limitation of this algorithm, however, is that it can be applied only when sequences are initiated by non-renewable events. We shall come back to this question later, in the context of an example.

Multiple Correspondence Analysis of Sequences

van der Heijden (1987, Ch. 8) illustrates and advocates the use of multiple correspondence analysis in the study of sequences⁵. This technique has become widespread in the analysis of qualitative data in the social sciences. The type of input data is similar to that which we discussed for binary sequences.

In the context of life course research, multiple correspondence analysis is useful for synthesizing the cross-sectional situation at each point in time, as well as for analysing the differences between individuals and identifying those individuals who are particularly “distant” from the mean. Graphical inspection is fundamental to this method. The applications that have been carried out up to now (which have generally focused on diaries and time-budgets) are substantially cross-sectional, and sometimes the time points need to be aggregated. Nevertheless, this technique can become particularly useful when sequences are generated by non-renewable events, since cross-sectional situations themselves depend on the past.

A Demographic Application

In the social sciences, most applications of sequence analysis have focused on the topic of working histories. There is still a lack of applications related to demographic trajectories such as union histories, childbearing, and residential mobility. We shall illustrate here one example. Billari and Piccarreta (2001) applied the monothetic divisive algorithm illustrated in the last section to study the transition to adulthood, using data from the Italian Fertility and Family Survey. The domains considered in their paper are education, the labour market, living arrangement, behaviour regarding sexual intercourse, union formation, and parenthood. Time is measured in years, and each variable is an indicator that represents whether one has experienced the marker event for each domain in a given year between the ages of 20 and 35 ($d = 6$ and $h = 16$, using the notation introduced in Section 2). The six variables that describe the state occupied in each domain at time t are:

- EDU_t = having finished education by the t -th year of age
- JOB_t = having entered the labour market by the t -th year of age
- LEA_t = having left the parental home by the t -th year of age
- SEX_t = having had sexual intercourse by the t -th year of age

- UNI_t = having entered a union by the t -th year of age
- CHI_t = having had a child by the t -th year of age.

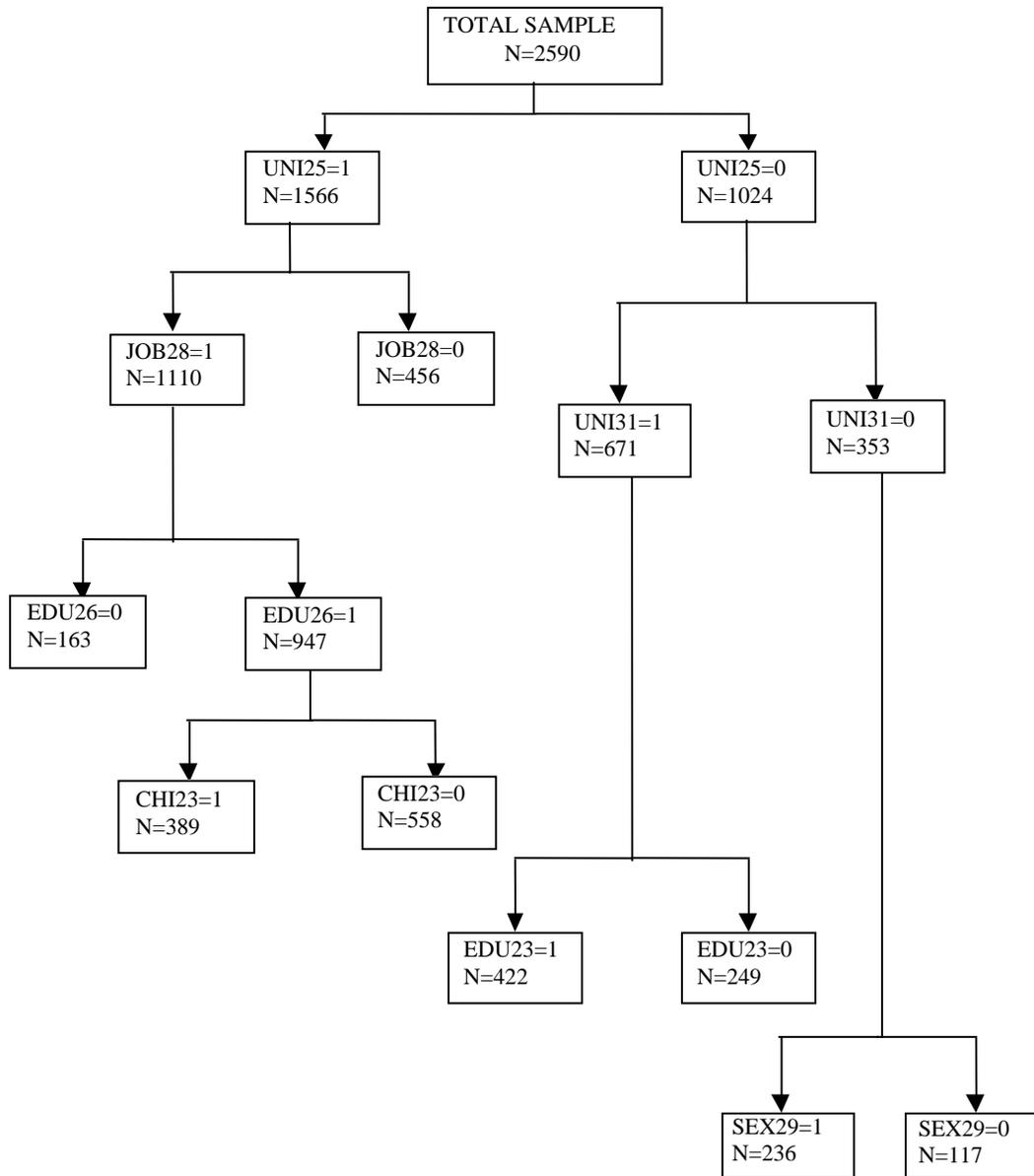
The property of sequences initiated by non-renewable events thus holds. Applying the divisive algorithm yields 8 groups (see Figure 1 for a tree representation of the splitting).

Our interest here is in seeing which variables are intervening at each stage of the split in order to be able to characterise the groups. At the first step, the splitting variable is UNI_{25} (having had a union by the age of 25). Hence, it seems that entering a union at a young age (relative to the local standards) is the most important indicator of differences between young Italians. This is consistent with certain views expressed in the literature that emphasised the role of unions in Southern European countries. Then, this ‘early union’ group is split according to JOB_{28} . In the case of late (or possibly no) labour force participation, this corresponds to a traditional female pattern (in fact, this group with early union and late or no job has the highest share of women among all the groups). The third step separates out those who have not entered a union before age 25 according to UNI_{31} ; again, the central role of marriage is being underlined here. The fourth step splits those entering an early union and having an early job according to EDU_{26} , then the group with a comparatively early end of schooling is split according to CHI_{23} . This leads to the formation of an ‘early-on-everything’ group and of two other groups with a comparatively late end of schooling and parenthood. At the sixth step, EDU_{23} splits the group of people entering a union between ages 25 and 31. The last step divides people not having entered a union at age 31 according to SEX_{29} ; this leads to the establishment of a group of people who have had no sexual relation before their thirtieth birthday.

Although it is not easy to interpret every single split, some features appear to be constant. Among young Italians, union formation plays a crucial role in classifying life courses in the transition to adulthood. Leaving the parental home is *not* important, which must certainly be connected with the high degree of synchronisation of this event with union formation. The transition to parenthood plays only a marginal role. Sexual behaviour matters only for people outside unions. As a consequence of having used the monothetic divisive algorithm, group membership is perfectly defined for specific life courses, and it becomes feasible to define groups of individuals according to the timing and sequencing of events in their transition to adulthood.

Billari and Piccarreta analysed also the basic demographic determinants (gender and cohort) of group membership. Group 1, with early union and with late (or no) labour force participation, largely corresponds to traditional absence of women from the labour force. There are only very few men in this group, and younger cohorts are underrepresented. Group 3 represents an average of sorts, without cohort and gender particularities, and it represents more than a fifth of

Figure 1.
Tree Representation of the Monothetic Divisive Algorithm Application



(source: Billari and Piccarreta, 2001).

the sample. Group 4 exhibits a faster transition to adulthood, and it is mainly composed of women; however, no cohort pattern is evident. Group 6, with unions between ages 25 and 31, has comparatively more men but no clear cohort dynamics. Group 8 has the highest share of “young” males; it comprises those who seem to be in line with the standard view on patterns of transitions to adulthood for more recent cohorts, which means that they have had sexual relationships but not within unions.

The main weakness of this approach is that it does not explicitly allow splitting by using the sequencing of events that belong to parallel careers. This is handled only in an indirect way, by focusing on the best possible split according to age. A suggestion to handle simultaneously the timing, sequence and the number of events in a holistic perspective can be seen in Billari et al. (2000).

Data Collection: Not Just Event Histories

What data collection procedures should be used if one wishes to construct a sequence representation of life courses? The answer is the same as for any longitudinal analysis: we need data that allow us to follow individuals over time. In this section, I briefly discuss this topic, and in the next section with ‘synthetic data.’

First, it is useful to note that, when the time scale for the available information is discrete, a sequence representation is equivalent to that of event histories (Rohwer and Pötter, 2000). I have already discussed this in the example seen in Section 2. Another example may be of some help. In a fictitious life course, there are four possible states denoted by: 0 (in school, not employed), 1 (out of school, not employed), 2 (in school, employed) and 3 (out of school, employed). An event history consists of only two events: an individual starts a job at month 5, and stops schooling at month 8. The monthly sequence can be derived from the event history as: 000002223333. Since the representation of the life course as an event history is equivalent to a representation of life course as a sequence using the same time scale, any instrument that allows the construction of event histories can also be used to produce sequences of states. This means that retrospective surveys can be, and in fact have been, used also to produce sequence data. It is not surprising that the technique of data collection known as the “life history calendar” (Freeman et al., 1988; Axinn et al., 1999) is based on the idea of representing life courses in a fashion similar to the sequence of states. Such methods are embedded in the broader spectrum of the collection of “life history matrices” (Settersten and Mayer, 1997) in life course research.

With retrospective data, we have the usual problems of having to rely on the memory of interviewees and of missing data. The problem of missing data can have different consequences in sequence analysis from it has for event history analysis. In fact, if we wish to study a life course in its entirety, the information

is incomplete even if only one of the events is wrongly dated. Possible strategies to deal with this problem include both the imputation techniques and defining an extra ‘missing state.’ In a retrospective survey, asking for the time order (sequencing) of events rather than a simple recording of events can serve as an advantage both in terms of quality check and performing analyses based on more reliable data.

The ideal source for sequence data is a population register, provided that it contains the information we are interested in. It has the advantage of providing the same data as a retrospective survey but without the problems of recall. In addition, a population register usually contains less missing information. The disadvantage is, however, twofold: population registers are rare, costly, and in some countries it is not even legal to keep registers; and, we can only construct sequences for trajectories that are officially recorded.

Nevertheless, less information is required for constructing a sequence representation than for constructing a full event history. The linkage of records from different censuses can provide a sequence that is sufficiently long and complex that specific techniques are required. For instance, if one links three census records with the present state (e.g., residential location) at each census and two recalled states for the pre-census period, one will have a sequence of 9 time points. Of course, the information about the intervals between the time points is lost in this case.

An additional and more diffused source of sequence data is provided by panel surveys. Such surveys normally collect the states occupied by individuals at various points in time. They do not necessarily provide full event histories, so discrete-time event history models have been used for the analysis of such data. Consequently, this type of source may prove to be rather powerful for the construction of sequences of states.

Synthetic Data and Sequence Analysis: The Case of Demographic Projections

Another type of data that might prove useful for sequence analyses consists of ‘artificially’ created life courses, known as ‘synthetic biographies’.⁶ Demographic projections based on microsimulation methods can actually produce such biographies, when several states (e.g. education, labour force status, living arrangement, family status, etc.) are taken into account⁷.

As Lutz (1997) states: “Computers can be also used to generate new virtual people. Such computer-generated individuals are only partial, inadequate images of real people. They may, nevertheless, be generated in a way that they carry some of the characteristics of real people that we consider decisive in determining their own behaviour and their impact on other people”.

Once the results of population projections through microsimulation techniques are available, we can further use the techniques of sequence analysis to a) check the results of projections for their consistency and sensitivity to different hypotheses; and, b) present the results of projections from a longitudinal perspective. The latter can be done by identifying individuals with specific life courses and/or grouping individuals according to their life courses by using the clustering algorithms developed for sequences.

Final Remarks

The representation of life courses as sequences and related analytical methods described in this paper can serve as powerful tools for studying the demographic components of life courses, as it has been in the study of work careers. We can expect that in the near future demographers will make a greater use of sequence analysis in their research work.

A few questions regarding the use of sequence analysis in demography still remain to be addressed (Wu, 2000). First, demographers rarely use microsimulation for projection purposes, and therefore there has been no effort until now to produce and study synthetic biographies, not to speak of applying sequence analysis to these biographies. Second, further research is needed to examine how the time scale used for sequence representations can or cannot distort the results. The 'ideal' time scale unquestionably depends on the topic under investigation and on the availability of data. Demographic data are generally available only on yearly basis, and event histories are often available as monthly data; but considerations of confidentiality may place restrictions on releasing such data. This is one of the impediments that researchers interested in doing sequence analysis may face everywhere. Third, how can one sensibly deal with right censoring, which is almost always present when analysing life courses⁸? Given that the length of sequences has itself an impact on the analyses, we need to clearly distinguish between censored sequences and short sequences. The problem then is how to perform a joint analysis using censored and non-censored sequences. Finally, it would be useful to connect the sequence analysis outlined in this paper to other more explanatory types of longitudinal analysis, as it has been done in genetics where, for example, sequences analysis and Markov models are respectively different tools for the analysis of outcomes and for investigating how such outcomes are generated (Waterman, 1995).

Appendix

Software for Sequence Analysis in the Social Sciences

There is an abundance of software for sequence analysis in the natural sciences (a review can be found in Abbott, 1997). However, some of the problems typical in the social sciences (shorter sequences but greater amounts of them) call for specific software. To date, two programs have been developed for the study of life courses represented as sequences. Fortunately, they are both available in the public domain.

Optimize was developed by Abbott (1997) and his colleagues. It is specially designed for optimal matching analysis. It has some limitations on the number of sequences (up to 150 sequences at a time).

TDA (Transition Data Analysis) is now a general program for statistics and data analysis. It was developed by Rohwer and Pötter (2000), and it provides a large number of commands for the complex description of sequences, as well as for the comparison of sequences (allowing for multiple sequences for each individual). Cluster analysis and correspondence analysis are also included in the more recent versions.

Acknowledgments:

Preliminary presentations based on this paper were given at the workshop “Projections of living arrangements, household and family structures”, Federal Institute for Population Research, Wiesbaden, August 1999 and at the Workshop on Longitudinal Research in Social Science - a Canadian Focus, University of Western Ontario, London, Ontario, Canada, October 1999. I gratefully acknowledge the participants at these workshops for their stimulating comments, Rajulton Fernando and Gert Hullen for comments and suggestions. Karl Brehmer has provided kind help in polishing language. The views expressed in this paper are the author’s own views and do not necessarily represent those of the Max Planck Institute for Demographic Research.

End Notes:

1. We do not discuss in this paper an important problem: the fact that macro-effects emerging from individual behaviours cannot be directly studied by analysing individual life histories only.
2. Some of the materials covered in Sections 2, 3, and 4 are based on past research conducted with other colleagues, mostly from Billari and Rohwer (1998) and Billari and Piccarreta (2000).

3. BioBrowser (ModGen Biography Browser) is a tool developed by Statistics Canada to supplement the ModGen language used for dynamic longitudinal microsimulation modelling. It allows the user to examine the life course of an actor and her/his close relatives graphically with different representations. Even if it is not based on discrete time, the representation produced can also be used in such framework. The software and the documentation can be downloaded at <http://www.statcan.ca/english/spsd/model.htm>.
4. This term does not necessarily correspond to the traditional notion of “cross-sectional” in demography.
5. In fact, van der Heijden refers to such data as ‘event history data’.
6. I borrow the term “synthetic biography” from Frans Willekens.
7. Macrosimulation methods for demographic projection, such as those based on cohort-component methods or more sophisticated multistate dynamics embedded in software like LIPRO (van Imhoff, 1994), provide pictures of the population at aggregate levels. They do not reflect the longitudinal evolution of (groups of) individuals nor do they produce individual synthetic biographies; hence, the sequence analysis cannot be used. Other methods that keep track of the evolution of individuals in a state space across time are also possible (e.g., more theoretically oriented agent-based models).
8. Some authors (e.g. Halpin and Chan, 1998) seem to argue that in principle, censoring should not be considered a problem when one is dealing with techniques that can handle sequences of different length. However, the conceptual difference between having sequences of different length because one's labour or union career is short, on the one hand, and because we do not know how long it is, on the other, cannot be disregarded.

References:

- Abbott A. 1995. "Sequence Analysis: New Methods for Old Ideas," *Annual Review of Sociology* 21: 93-113.
- Abbott A. 1997. *Program Optimize*. Chicago: University of Chicago.
- Abbott A. and A. Tsay. 2000. "Sequence Analysis and Optimal Matching Methods in Sociology," *Sociological Methods & Research* 29: 3-33.
- Axinn W. G., L. D. Pearce and D. Ghimire. 1999. "Innovations in Life History Calendar Applications," *Social Science Research* 28: 243-264.

- Billari F.C. and G. Rohwer. 1999. "Some developments in methods for life course analysis and their applications," in: *Atti della XXXIX Riunione Scientifica della Società Italiana di Statistica, Comunicazioni invitate*. Napoli: Istituto Universitario Navale: 323-334.
- Billari F.C. and R. Piccarreta. 2001. Life Courses as Sequences: an Experiment in Clustering via Monothetic Divisive Algorithms, in S. Borra, R. Rocci, M. Vichi and M. Schader (Eds.). *Advances in Classification and Data Analysis*. Heidelberg: Springer.
- Billari F.C. and R. Piccarreta. 2000. "Studying demographic life courses with sequence analysis," *mimeo*. Rostock: Max Planck Institute for Demographic Research.
- Billari F.C., J. Fürnkranz and A. Prskawetz. 2000. "Timing, Sequencing and Quantum of Life Course Events: a Machine Learning Approach," WP 2000-010. Rostock: Max Planck Institute for Demographic Research.
- Blossfeld H.-P. and G. Rohwer. 1995. *Techniques of Event History Modeling. New Approaches to Causal Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Camerer C. 1995. Individual Decision Making. In Kagel J.H. and A.E. Roth (Eds.). *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.
- Chan T.W., 1999, "Optimal Matching Analysis," *Social Research Update*, 24.
- Deaton A. and J. Muellbauer, 1980. *Economics and consumer behavior*. Cambridge, UK: Cambridge University Press.
- Freeman D., A. Thornton, D. Camburn, D. Alwin and L. Young-DeMarco. 1988. "The Life History Calendar: A Technique for Collecting Retrospective Data," *Sociological Methodology*: 37-68.
- Giddens A.. 1991. *Modernity and self-identity: Self and society in the late modern age*. Cambridge, UK: Polity Press.
- Giele J.Z. and G.H. Elder Jr. 1998. Life Course Research. Development of a Field. in Giele J.Z. and G.H. Elder Jr. (Eds.). *Methods of Life Course Research. Qualitative and Quantitative Approaches*. Thousand Oaks, California: Sage.

- Halpin B. and T. W. Chan. 1998. "Class careers as sequences: an optimal matching analysis of work-life histories," *European Sociological Review* 14: 111-130.
- Heckhausen J. 1999. *Developmental Regulation in Adulthood. Age-Normative and Sociostructural Constraints as Adaptive Challenges*. Cambridge, UK: Cambridge University Press.
- Lutz W. 1997. Introduction: Purpose of FAMSIM. In Lutz W. (Ed.). *FAMSIM-Austria. Feasibility Study for a Dynamic Microsimulation Model for Projections and the Evaluation of Family Policies Based on the European Family and Fertility Survey*. Vienna: Austrian Institute for Family Studies.
- Mayer K.U. and N.B. Tuma (Eds.). 1990. *Event history analysis in life course research*. Madison, WI: University of Wisconsin Press.
- Myers E. W. 1995. Seeing Conserved Signals: Using Algorithms to Detect Similarities Between Biosequences, in Lander E.S. and M.S. Waterman (Eds.). *Calculating the Secrets of Life*. Washington, DC: National Academy Press.
- Rohwer G. 1994. *Kontingente Lebensverläufe. Soziologische und statistische Aspekte ihrer Beschreibung und Erklärung*. Bremen: Universität Bremen.
- Rohwer G. and U. Pötter. 2000. *TDA User's manual*, Bochum: Ruhr-Universität Bochum.
- Rohwer G. and H. Trappe. 1999. Possibilities and difficulties in life course description, in W. Voges (Ed.). *Dynamic approaches to comparative social research. Recent developments and applications*. Aldershot: Avebury Publishers, pp. 146-167.
- Sankoff D. and J. B. Kruskal (Eds.). 1983. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Reading, MA: Addison-Wesley.
- Settersten R. A. and K. U. Mayer. 1997. "The Measurement of Age, Age Structuring and the Life Course," *Annual Review of Sociology* 23, 233-261.
- Statistics Canada. 1999. *BioBrowser: The ModGen Biography Browser. Users Guide. Version 3.1*, Ottawa: Statistics Canada.

- van der Heijden P. G. M. 1987. *Correspondence analysis of longitudinal categorical data*. Leiden: DSWO Press.
- van Imhoff E. 1994. *LIPRO 3.0 User's Guide*. Working Paper 1994/1b. The Hague: NIDI.
- Waterman, M. S. 1995. *Introduction to Computational Biology. Maps, sequences and genomes*. London and New York: Chapman & Hall.
- Wehner S. 1999. *Exploring and visualizing event history data*. Materialien aus der Bildungsforschung Nr. 65. Berlin: Max-Planck-Institut für Bildungsforschung.
- Wu L. L. 2000. "Some Comments on 'Sequence Analysis and Optimal Matching Methods in Sociology': Review and Prospect," *Sociological Methods & Research* 29: 41-64.
- Wunsch G. and M. Termote. 1978. *Introduction to Demographic Analysis. Principles and Methods*. New York and London: Plenum Press.
- Yamaguchi K. 1991. *Event History Analysis*. Newbury Park, California: Sage.